SEMA SIMAI Springer series 28

María Luz Muñoz-Ruiz
Carlos Parés
Giovanni Russo *Eds.*

Recent Advances in Numerical Methods for Hyperbolic PDE Systems

NumHyp 2019

SEMA SIMAI SOCIETÀ ITALIANA DI MATEMATICA APPLICATA E INDUSTRIALE
Springer

SEMA SIMAI Springer Series

Volume 28

As of 2013, the SIMAI Springer Series opens to SEMA in order to publish a joint series aiming to publish advanced textbooks, research-level monographs and collected works that focus on applications of mathematics to social and industrial problems, including biology, medicine, engineering, environment and finance. Mathematical and numerical modeling is playing a crucial role in the solution of the complex and interrelated problems faced nowadays not only by researchers operating in the field of basic sciences, but also in more directly applied and industrial sectors. This series is meant to host selected contributions focusing on the relevance of mathematics in real life applications and to provide useful reference material to students, academic and industrial researchers at an international level. Interdisciplinary contributions, showing a fruitful collaboration of mathematicians with researchers of other fields to address complex applications, are welcomed in this series.

THE SERIES IS INDEXED IN SCOPUS

More information about this series at http://www.springer.com/series/10532

María Luz Muñoz-Ruiz ·
Carlos Parés · Giovanni Russo
Editors

Recent Advances in Numerical Methods for Hyperbolic PDE Systems

NumHyp 2019

Springer

Editors
María Luz Muñoz-Ruiz
Department of Mathematical Analysis,
Statistics and Applied Mathematics
University of Málaga
Málaga, Spain

Carlos Parés
Department of Mathematical Analysis,
Statistics and Applied Mathematics
University of Málaga
Málaga, Spain

Giovanni Russo
Department of Mathematics
and Computer Science
University of Catania
Catania, Italy

ISSN 2199-3041 ISSN 2199-305X (electronic)
SEMA SIMAI Springer Series
ISBN 978-3-030-72849-6 ISBN 978-3-030-72850-2 (eBook)
https://doi.org/10.1007/978-3-030-72850-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The present volume contains selected papers issuing from the sixth edition of the International Conference Numerical Methods for hyperbolic problems that took place in Málaga, Spain, from 17 to 21 June 2019. Málaga was a sub-venue of the 2019 International Congress on Industrial and Applied Mathematics held in Valencia (Spain) from 15 to 19 July 2019, and NumHyp2019 was considered as one of the satellite events of ICIAM2019. Moreover, NumHyp2019 was a key activity of the European innovative training network entitled Modelling and Computation of Shocks and Interfaces (ModCompShock). The conference took place in the building of the Instituto de Estudios Portuarios located in the port of Málaga, close to the heart of the city.

NumHyp is a series of biannual conferences that began with a meeting in Castro Urdiales, Spain, in 2009. Further editions of this conference were held in Roscoff, France, in 2011, Aachen, Germany, in 2013, Cortona, Italy, in 2015, and Monte Verità, Switzerland, in 2017. These conferences focus on recent developments and new directions in the area of numerical methods for hyperbolic and convection-dominated partial differential equations (PDEs). These PDEs arise in a large number of models in physics and engineering. Prominent examples include compressible and incompressible Euler and Navier–Stokes equations, shallow water equations, magnetohydrodynamics, multiphase fluid models, etc. Examples of application areas are aerodynamics, oceanography, plasma physics, solid mechanics, geophysics, environmental sciences, etc.

These PDEs have been subject of extensive analytical and numerical investigation over the last decades. It is widely known that their solutions can exhibit very complex behaviour including formation and propagation of singularities such as shock waves, sensitive dependence to initial conditions, presence of multiple spatio-temporal scales, and appearance of turbulent regimes. The design and the analysis of numerical methods with good properties to solve them are still a challenge.

More than 70 participants coming from 12 different countries attended the event, and 10 invited talks, 30 oral contributions, and 14 posters were presented. More details can be found at the conference website: http://eventos.uma.es/go/NumHyp19.

This special issue contains 11 chapters selected from among the invited lectures and contributed talks covering several state-of-the-art numerical techniques and/or applications of hyperbolic systems. They have been organized in three parts:

1. Numerical methods for general problems. The chapters in this part put the emphasis on the design of new schemes useful for general families of hyperbolic problems. More precisely, the following aspects are discussed:

   – Incomplete Riemann solvers for conservative and nonconservative hyperbolic systems and Jacobian-free methods (Chapter 1).
   – Entropy-based methods for uncertainty quantification (Chapter 2).
   – High-order well-balanced methods for systems of balance laws (Chapter 3).
   – IMEX methods for multi-scale scalar conservation laws (Chapter 4).

   The application of the methods discussed in the different chapters is illustrated in a number of flow models, such as Euler equations, ideal magnetohydrodynamics, and one-layer and multi-layer shallow water models.
2. Numerical methods for specifc problems: The chapters in this part focus on the design of numerical methods with good properties for specific flow models. More precisely,

   – A staggered pressure correction numerical method for solving a model of turbulent deflagrations in industrial applications (Chapter 5).
   – Invariant domain preserving finite volume methods for the compressible Euler and Navier–Stokes equations (Chapter 6).
   – A level-set ghost-fluid high-order DG method for compressible multiphase flow models (Chapter 7).
   – All Mach number entropy stable methods for the compressible Euler equations (Chapter 8).
   – Residual-based methods for sediment-transport models (Chapter 9).

3. New flow models: Chapters 10 and 11 put the emphasis on the derivation of new nonlinear dispersive shallow water models.

Chapters 1, 2, 5, 6, 7, and 10 present the contents of the invited talks given by J.M. Gallardo, M. Frank, R. Herbin, H. Mizerová, C.D. Munz, and J. Sainte-Marie, respectively.

We would like to address our warmest thanks and gratitude to all who have made this book possible: first of all, to all the speakers of NumHyp2019 for their valuable contributions and, very specially, to those who accepted our invitation to contribute to this volume and next to the anonymous referees that helped the authors to improve the quality of their manuscripts. We would also like to thank the members of the scientific committee for their support and help in the speakers

selection and to those of the organizing committee for taking care of the organization of the event. We would like to thank the sponsors without whom NumHyp2019 would not have been possible: in addition to the already mentioned ITN ModCompShock, we are really grateful to the University of Málaga and the Sociedad Española de Matemática Aplicada (SEMA). Many thanks to the organizing committee of ICIAM 2019 for having considered this conference as a satellite event. We also thank the Springer staff for their help and support during the edition process and very specially to Francesca Bonadei, executive editor in charge. Finally, we thank the editorial board of the SEMA/SIMAI Springer series for having accepted this volume and the editor in charge, Paolo Zunino, for his helpful comments.

M. L. Muñoz-Ruiz
C. Parés
G. Russo

Contents

Numerical Methods for General Problems

Incomplete Riemann Solvers Based on Functional Approximations to the Absolute Value Function

José M. Gallardo, Manuel J. Castro, and Antonio Marquina

Abstract We give an overview on the work developed in recent years about certain classes of incomplete Riemann solvers for hyperbolic systems. These solvers are based on polynomial or rational approximations to $|x|$, and they do not require the knowledge of the complete eigenstructure of the system, but only a bound on the maximum wave speed. Our solvers can be readily applied to nonconservative hyperbolic systems, by following the theory of path-conservative schemes. In particular, this allows for an automatic treatment of source or coupling terms in systems of balance laws. The properties of our schemes have been tested with some challenging numerical experiments involving systems such as the Euler equations, ideal magnetohydrodynamics equations and multilayer shallow water equations.

1 Introduction

Since the early work of Godunov [19], Riemann solvers constitute a fundamental ingredient in the design of robust and accurate numerical methods for hyperbolic conservation laws. Usually, Riemann solvers can be classified as *complete* or *incomplete*, depending if all the characteristic waves in the solution of the exact Riemann problem are considered or not. Among the class of complete Riemann solvers, Roe's method [29] is one of the most widely used, as it usually provides the best resolution of the Riemann wave fan. However, when analytic expressions for the eigenstructure of the system are not available or they are difficult to compute, Roe's method may

J. M. Gallardo (✉) · M. J. Castro
Department of Análisis Matemático, Estadística e I. O., y Matemática Aplicada, University of Málaga, Málaga, Spain
e-mail: jmgallardo@uma.es

M. J. Castro
e-mail: mjcastro@uma.es

A. Marquina
Department of Matemática Aplicada, University of Valencia, Valencia, Spain
e-mail: marquina@uv.es

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*, SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2_1

be computationally expensive. Therefore, in certain situations it is preferable to consider incomplete Riemann solvers, for which only part of the spectral information is needed. In these cases, an important drawback may be the lack of resolution of internal waves in complex scenarios.

The numerical diffusion of a given numerical flux is determined by its *viscosity matrix*. In the case of Roe's method the viscosity matrix is $|A|$, the absolute value of the Roe matrix of the system, which may be difficult to compute as it requires the knowledge of the complete eigenstructure of A. A number of incomplete Riemann solvers based on appropriate approximations to $|A|$ have been proposed in the literature. One of the earliest examples is given by the local Lax-Friedrichs (or Rusanov) method, in which $|A|$ is approximated using only the largest eigenvalue of the system. Another very popular approach is the HLL method [20], where $|A|$ is approximated be means of a linear polynomial evaluation $P(A)$, where $P(x)$ interpolates $|x|$ at the smallest and largest eigenvalues of A. On the other hand, the paper [14] contains the first construction of a simple approximation to $|A|$ by means of a polynomial that approximates $|x|$ without interpolating it exactly on the eigenvalues.

The latter approach is the basis of the general framework proposed in [8], where PVM (Polynomial Viscosity Matrix) methods were introduced. The viscosity matrix of a PVM method is built as a polynomial evaluation $P(A)$ of the Roe matrix or the Jacobian of the flux at some other average value. It is worth noticing that a number of well-known methods in the literature can be viewed as particular cases of PVM schemes: Lax-Friedrichs, Rusanov, HLL, FORCE, Roe, etc. (see also [13, 23, 31]). An additional feature of PVM methods is that they can be defined in the general framework of nonconservative hyperbolic systems, which allows to construct natural extensions of the standard schemes cited before for solving problems in nonconservative form.

To ensure the stability of a PVM method, the graph of the basis polynomial $P(x)$ must be over the graph of the absolute value function. On the other hand, as $P(x)$ is closer to $|x|$ in the uniform norm, the behavior of the associated PVM method will be closer to that of Roe's method. It follows then that it is possible to use accurate approximations to $|x|$ for building PVM schemes resembling Roe's method, but with a much smaller computational cost. Following this idea, a PVM scheme based on Chebyshev polynomials, which provide optimal uniform approximations to $|x|$, was proposed in [10]. This idea was further extended in the same paper to the case of rational functions, which greatly improve the order of approximation to $|x|$. The resulting schemes were denoted as RVM (Rational Viscosity Matrix). In fact, RVM schemes based on Newman [24] approximations provided similar performances as Roe's method, but with a much smaller computational cost. As the only difference between PVM and RVM methods rely on the kind of basis function chosen, in this work we will use the term AVM (Approximate Viscosity Matrix) to refer to both of them. We remark that AVM methods constitute a class of general-purpose Riemann solvers, which are constructed using only an estimate of the spectral radius of the Roe matrix or the Jacobian of the system evaluated at an average state. As an additional advantage, unlike Roe's method, no entropy-fix is needed in the presence of sonic points, as long as the basis function does not cross the origin. Recently, a

fully two-dimensional version of AVM schemes has been proposed in [18], where multidimensional effects are taken into account through the approximate solution of two-dimensional Riemann problems.

The Osher-Solomon (OS) scheme [26] is a nonlinear and complete Riemann solver which enjoys a number of interesting features: it is robust, smooth, entropy-satisfying, and good behaved when computing slowly-moving shocks. As a drawback, its practical implementation is complex and computationally expensive, as it requires the computation of a path-dependent integral in phase space (see [30]). For this reason, its practical application has been restricted to certain systems, e.g., the compressible Euler equations. In [15, 16], the authors proposed a variant of the OS method combining linear paths and a Gauss-Legendre quadrature formula. This led to a simplified version of the OS scheme, denoted as DOT (Dumbser-Osher-Toro), which conserves its good properties and it is applicable to general hyperbolic systems. In particular, the viscosity matrix of a DOT solver is defined as a linear combination of the absolute value matrix of the Jacobian of the physical flux evaluated at certain quadrature points. As the practical computation of these matrices can be expensive, they could be approximated in an efficient way following the same technique behind AVM methods. This idea was explored in [11], leading to the class of AVM-DOT solvers. In particular, it was shown that Chebyshev-based AVM-DOT solvers admit a Jacobian-free implementation, in which only evaluations of the physical flux are needed. This kind of methods is particularly interesting when solving systems in which the Jacobian involves complex expressions: see [12].

Both classes of AVM and AVM-DOT solvers can be extended to the case of nonconservative hyperbolic systems, following the theory of path-conservative schemes [28]. In particular, this includes the important case of hyperbolic systems of conservation laws with source terms and nonconservative products. In the conservative case, the proposed schemes have been applied to a number of challenging problems in ideal gas dynamics, magnetohydrodynamics (MHD) and relativistic MHD (RMHD). Multilayer shallow water systems have been considered as a representative example in the nonconservative framework, as they include both source and nonconservative coupling terms [10–12]. In all the cases, the numerical tests indicate that the proposed schemes are robust, running stable and accurate with a satisfactory time step restriction.

2 Approximate Viscosity Matrix (AVM) Methods

In this section we give an overview of PVM [8] and related methods developed in recent years [10, 12]. For the sake of clarity, we first focus on the case of a system of conservation laws. Extensions to the nonconservative case will be treated later in Sect. 4.

Let us consider a system of conservation laws

$$\partial_t w + \partial_x F(w) = 0, \tag{1}$$

where $w(x,t)$ takes values on an open convex set $\mathcal{O} \subset \mathbb{R}^N$ and $F\colon \mathcal{O} \to \mathbb{R}^N$ is a smooth flux function. The numerical solution of the Cauchy problem for (1) is computed by means of a finite volume method of the form

$$w_i^{n+1} = w_i^n - \frac{\Delta t}{\Delta x}(F_{i+1/2} - F_{i-1/2}), \tag{2}$$

where w_i^n is an approximation to the average of the exact solution at the cell $I_i = [x_{i-1/2}, x_{i+1/2}]$ at time $t^n = n\Delta t$ (the dependence on time will be dropped unless necessary). The numerical flux is assumed to be written as

$$F_{i+1/2} = \frac{F(w_i) + F(w_{i+1})}{2} - \frac{1}{2} Q_{i+1/2}(w_{i+1} - w_i), \tag{3}$$

where the *viscosity matrix* $Q_{i+1/2}$ controls the numerical diffusion of the scheme.

We will assume that system (1) is *hyperbolic*, i.e., the Jacobian matrix of the flux at each state $w \in \mathcal{O}$,

$$A(w) = \frac{\partial F}{\partial w}(w),$$

can be diagonalized as

$$A = PDP^{-1},$$

where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$, λ_i are the eigenvalues of A, and the columns of the matrix P are the associated right eigenvalues of A. As it is usual, we denote the positive and negative parts of A, respectively, as

$$A^+ = PD^+P^{-1}, \quad A^- = PD^-P^{-1},$$

where $D^\pm = \mathrm{diag}(\lambda_1^\pm, \ldots, \lambda_N^\pm)$, with $\lambda_i^+ = \max(\lambda_i, 0)$ and $\lambda_i^- = \min(\lambda_i, 0)$. It is clear that $A = A^+ + A^-$. The absolute value of A is then defined as

$$|A| = A^+ - A^-.$$

One of the most widely used Riemann solvers for (1) was proposed by Roe in [29]. It usually provides the best resolution of the Riemann wave fan, although for complex systems the method can be computationally expensive. This is due to the fact that Roe's method is a *complete* Riemann solver, in the sense that it uses all the eigenstructure of the system. Therefore, for complex systems, or systems for which the eigenstructure is not known, *incomplete* Riemann solvers may be preferred: they use few characteristic information and are thus easier to implement and computationally efficient.

It is important to note that Roe's method can be written in the form (3) with viscosity matrix $Q_{i+1/2} = |A_{i+1/2}|$, where $A_{i+1/2}$ is a Roe matrix for the system. Several numerical methods have been developed by using approximations to $|A_{i+1/2}|$ as viscosity matrices; see, e.g., [13, 14, 20, 30, 31] and the references therein.

PVM Methods

The original idea of PVM (Polynomial Viscosity Matrix) Riemann solvers [8] was based in approximating $|A_{i+1/2}|$ using an appropriate polynomial evaluation of $A_{i+1/2}$. Assume that $P(x)$ is a polynomial approximation of $|x|$ in the interval $[-1, 1]$, and let $\lambda_{i+1/2,\max}$ be the eigenvalue of $A_{i+1/2}$ with maximum modulus (or an upper bound of it). The numerical flux of the PVM method associated to $P(x)$ is given by (3) with viscosity matrix

$$Q_{i+1/2} = |\lambda_{i+1/2,\max}| P(|\lambda_{i+1/2,\max}|^{-1} A_{i+1/2}),$$

which provides an approximation to $|A_{i+1/2}|$, the viscosity matrix of Roe's method. Moreover, note that the best $P(x)$ approaches $|x|$, the closer the behavior of the associated PVM scheme will be to that of Roe's method. It is worth noticing that no spectral decomposition of the matrix $A_{i+1/2}$ is needed to build a PVM method, but only a bound on its spectral radius. This fact makes PVM methods greatly efficient and applicable to systems in which the eigenstructure is not known or difficult to obtain. In those cases in which a Roe matrix is not available or is difficult to compute, $A_{i+1/2}$ can be defined as the Jacobian evaluated at some average state.

Several well-known schemes in the literature can be interpreted as PVM methods, for example:

- Lax-Friedrichs: $P(x) = \frac{\Delta x}{\Delta t}$.
- HLL: $P(x) = \alpha_0 + \alpha_1 x$, where $P(S_L) = |S_L|$ and $P(S_R) = |S_R|$, being S_L and S_R approximations to the minimal and maximal speeds of propagation.
- Roe: In this case, $P(x)$ is the Lagrange polinomial which interpolates the set of points $(\lambda^{(j)}_{i+1/2}, |\lambda^{(j)}_{i+1/2}|)$, where $\lambda^{(j)}_{i+1/2}$ are the eigenvalues of the Roe matrix $A_{i+1/2}$.

Other examples include Rusanov, FORCE or Lax-Wendroff methods (see [8]). Another example of PVM method is the one proposed in [14], which constitutes one of the first attempts to construct a simple approximation of $|A|$ by means of a polynomial that approximates $|x|$ without interpolating it exactly on the eigenvalues.

The stability of a PVM scheme relies on the properties of the basis polynomial $P(x)$. In particular, the following *stability condition* should be verified:

$$|x| \leq P(x) \leq 1, \quad \forall\, x \in [-1, 1]. \tag{4}$$

It was proven in [8] that condition (4) implies that the associated PVM scheme is linearly L^∞-stable under a standard CFL restriction.

A well-known drawback of Roe's method is the need of an entropy fix to handle sonic flow properly, in order to avoid entropy-violating solutions. In PVM-type schemes there is no need of entropy fix as long as $P(0) \neq 0$.

In [10] we proposed a new class of PVM schemes based on *Chebyshev polynomials*, which provide optimal uniform approximations to the absolute value function. The Chebyshev polynomials of even degree $T_{2k}(x)$ are recursively defined as

$$T_0(x) = 1, \quad T_2(x) = 2x^2 - 1, \quad T_{2k}(x) = 2T_2(x)T_{2k-2}(x) - T_{2k-4}(x).$$

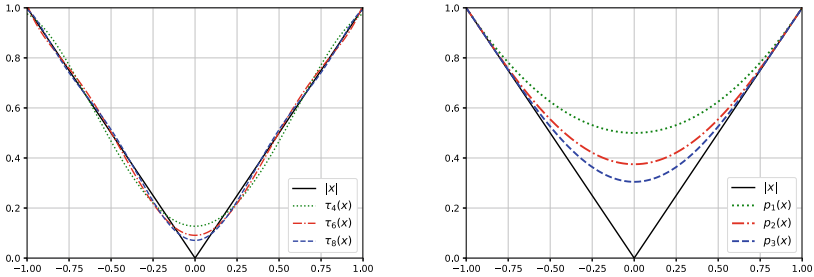


Fig. 1 Left: Chebyshev approximations $\tau_{2p}(x)$ for $p = 2, 3, 4$. Right: Internal polynomial approximations (6)

Then, for $p \geq 1$ we consider the polynomial of degree $2p$ given by (see Fig. 1, left)

$$\tau_{2p}(x) = \frac{2}{\pi} + \sum_{k=1}^{p} \frac{4}{\pi} \frac{(-1)^{k+1}}{(2k-1)(2k+1)} T_{2k}(x), \quad x \in [-1, 1], \tag{5}$$

which follows after truncation of the series expansion of $|x|$ in terms of Chebyshev polynomials. It is a classical result [2] that the order of approximation of $\tau_{2p}(x)$ to $|x|$ is optimal in the $L^{\infty}(-1, 1)$ norm. Moreover, the recursive definition of the polynomials $T_{2k}(x)$ provides an explicit and efficient way to compute $\tau_{2p}(x)$ (see the Appendix in [10]).

Notice that $\tau_{2p}(x)$ does not verify the stability condition (4) strictly: see Fig. 1 (left), where $\tau_{2p}(x)$ has been drawn for $p = 2, 3, 4$. This drawback was partially fixed in [10] in a rough manner, by substiting $\tau_{2p}(x)$ by $\tau_{2p}^{\varepsilon} = \tau_{2p}(x) + \varepsilon$, where ε is chosen as the minimum value such that $\tau_{2p}^{\varepsilon}(x)$ fulfills condition (4). However, this could cause incorrect approximations of the external waves.

In [12] we proposed another family of polynomials which approximate $|x|$ in a more elegant way, satisfying the stability condition (4) by construction. This class of *internal* polynomials are iteratively defined as follows (see Fig. 1 (right)):

$$p_0(x) \equiv 1, \quad p_{n+1}(x) = \frac{1}{2}\big(2p_n(x) - p_n(x)^2 + x^2\big), \quad n = 0, 1, 2, \ldots \tag{6}$$

RVM Methods

As it is well-known [24], rational functions provide much better approximations to $|x|$ than polynomials. This fact was used in [10] to create new families of very precise incomplete Riemann solvers, denoted as RVM (Rational Viscosity Matrix), following the same idea of PVM methods but using a rational function as basis instead of a polynomial.

In particular, two families of RVM methods will be considered here. The first one corresponds to *Newman* approximations, which are constructed as follows. For a given $r \geq 4$, consider a set of distinct points in $(0, 1]$, $X = \{0 < x_1 < \cdots < x_r \leq 1\}$, and build the polynomial

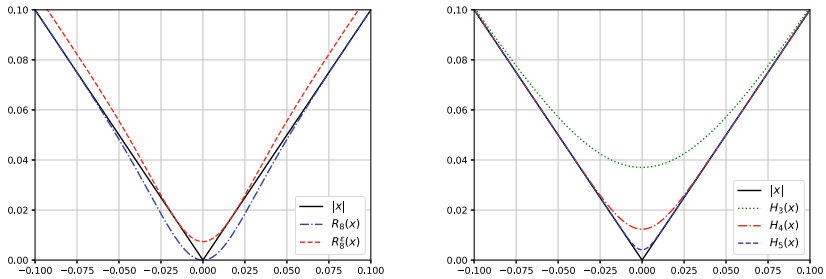


Fig. 2 Left: Comparison between $R_8(x)$ and $R_8^\varepsilon(x)$ for $x \in [-0.1, 0.1]$, with $\varepsilon \approx 7.37e-3$. Right: Halley rational approximations $H_r(x)$ in the interval $[-0.1, 0.1]$, for $r = 3, 4, 5$

$$p(x) = \prod_{k=1}^{r} (x + x_k).$$

The *Newman rational function* associated to the set X is then defined by

$$R_r(x) = x \, \frac{p(x) - p(-x)}{p(x) + p(-x)}.$$

It is easy to see that $R_r(x)$ interpolates $|x|$ at the points $\{-x_r, \ldots, -x_1, 0, x_1, \ldots, x_r\}$. Also notice that for even r both the numerator and denominator of $R_r(x)$ are of degree r. The uniform rate of approximation of $R_r(x)$ to $|x|$ depends on the choice of the set of nodes X. Several choices are possible (see [10]); here, we have considered Newman's original definition, which is given by $x_k = \xi^k$, with $\xi = \exp(-r^{-1/2})$.

Notice that the stability condition (4) is not fulfilled in any case, so a *modified approximation* of the form $R_r^\varepsilon(x) = R_r(x) + \varepsilon$ should be considered, as in the case of Chebyshev polynomials. A comparison between $R_r(x)$ and $R_r^\varepsilon(x)$ can be seen in Fig. 2 (left). The differences between using $R_r(x)$ or $R_r^\varepsilon(x)$ are particularly noticeable in the presence of sonic points: in this case, $R_r^\varepsilon(x)$ must be used to avoid entropy-violating solutions.

The second family of rational functions is based on iterative approximations. Note that the absolute value $|\bar{x}|$ of a given point $\bar{x} \in [-1, 1]$ can be viewed as the positive root of $f(x) = x^2 - \bar{x}^2$. It is then possible to approximate $|\bar{x}|$ using a root-finding algorithm, such as Newton's method, or the more precise choice given by the cubic Halley's method:

$$x_{k+1} = x_k \, \frac{x_k^2 + 3\bar{x}^2}{3x_k^2 + \bar{x}^2}.$$

Taking $x_0 = 1$ as initial guess, Halley's method is well-defined and converges to $\bar{x}$ (see [5]). The *Halley rational approximations* to $|x|$ are thus recursively defined as

$$H_0(x) \equiv 1, \quad H_{r+1}(x) = H_r(x) \, \frac{H_r(x)^2 + 3x^2}{3H_r(x)^2 + x^2}. \tag{7}$$

Notice that the degrees of the numerator and denominator of $H_r(x)$ are both equal to $3^r - 1$. It can be easily verified that $H_r(x)$ satisfies the stability condition (4) without further modifications. Figure 2 (right) shows the functions $H_r(x)$ for $r = 3, 4, 5$.

As it was commented before, another possibility is to use Newton's method instead of Halley's method. However, numerical experiments show that RVM-Halley methods provide much better resolution of internal waves than RVM-Newton schemes with a comparable computational cost.

AVM Methods

As it can be seen in the preceding paragraphs, the idea behind PVM and RVM methods is esentially the same, the difference depending only on using polynomials or rational functions to approximate the absolute value function. For this reason, we will encompass both kind of methods under the global name of AVM (Approximate Viscosity Matrix) methods.

Therefore, an AVM method is a finite volume method of the form (2), where the numerical flux is given by (3) with viscosity matrix

$$Q_{i+1/2} = f(A_{i+1/2}),$$

where $f : \mathbb{R} \to \mathbb{R}$ is a given function and $A_{i+1/2}$ is a Roe matrix or the Jacobian of the flux evaluated at some average state. The function f must verify some conditions:

- $f(x)$ is nonnegative and smooth.
- $f(A_{i+1/2})$ should be *easy* to evaluate; in particular, no spectral decomposition of $A_{i+1/2}$ should be needed, but only a bound on its spectral radius.
- L^∞-linear stability, which is accomplished under the condition

$$|x| \leq f(x) \leq \mathrm{CFL}\frac{\Delta x}{\Delta t}, \quad \forall\, x \in [\lambda^{(1)}_{i+1/2}, \lambda^{(N)}_{i+1/2}],$$

  where $\lambda^{(1)}_{i+1/2} \leq \cdots \leq \lambda^{(N)}_{i+1/2}$ are the eigenvalues of $A_{i+1/2}$.
- The graph of $f(x)$ should be as close as possible to the graph of $|x|$.

We end this section with some remarks regarding computational efficiency. The implementation of rational-based methods involves the computation of matrix powers and matrix inversions, while Chebyshev-based or internal-based methods can be expressed in such a way that only vector operations are involved. This makes these polynomial methods computationally cheaper than rational ones but, on the other hand, rational methods are far more precise. Thus, a compromise has to be accomplished between accuracy and computational cost. In general, it seems that for problems with solutions containing very complex patterns, rational-based methods perform better than polynomial ones in terms of the ratio between CPU times and computed errors. Otherwise, for mildly complex problems polynomial methods may be the preferred option.

An additional advantage of Chebyshev-based and internal-based methods is that they admit a Jacobian-free implementation (see Sect. 3), which is not possible for rational-based methods. This means that the numerical flux can be constructed using

only evaluations of the physical flux F at different states, thus avoiding the computation of Jacobian matrices. This point is particularly interesting for systems with complex physical fluxes (as, for example, the equations of RMHD), for which the calculation of the corresponding Jacobian may be a difficult or costly task.

3 Approximate DOT Solvers

The so-called *approximate DOT* (Dumbser-Osher-Toro) solvers, introduced in [11], combine the AVM technique with the universal Osher-type solvers proposed in [15]. These methods, that will be denoted in what follows as AVM-DOT, constitute simple and efficient approximations to the classical Osher-Solomon method [26], enjoying most of its interesting features and being applicable to general hyperbolic systems.

The numerical flux of the original Osher-Solomon method is given by

$$F_{i+1/2} = \frac{F(w_i) + F(w_{i+1})}{2} - \frac{1}{2}\int_0^1 \left|A(\Phi(s))\right|\Phi'(s)ds, \tag{8}$$

where $A(w)$ represents the Jacobian of the physical flux F evaluated at the state w, and Φ is a path in phase-space linking the states w_i and w_{i+1}. The path Φ for a DOT solver [15] is taken as the segment linking w_i and w_{i+1}, and the resulting integral is approximated using a Gauss-Legendre quadrature formula. Thus, the resulting DOT flux adopts the form (3) with viscosity matrix

$$Q_{i+1/2} = \sum_{k=1}^{q} \omega_k \left|A(w_i + s_k(w_{i+1} - w_i))\right|,$$

where ω_k and s_k are the weights and nodes of the quadrature formula. Now, the absolute value of the intermediate matrices could be approximated by using the technique of AVM methods. The resulting AVM-DOT solver associated to a function $f(x)$ has then the following form:

$$Q_{i+1/2} = \sum_{k=1}^{q} \omega_k \widetilde{P}^{(k)}_{i+1/2},$$

where

$$\widetilde{P}^{(k)}_{i+1/2} = \left|\lambda^{(k)}_{i+1/2,\max}\right| f\left(\left|\lambda^{(k)}_{i+1/2,\max}\right|^{-1} A^{(k)}_{i+1/2}, \right. \tag{9}$$

for $k = 1, \ldots, q$. In the above expression, $\lambda^{(k)}_{i+1/2,\max}$ denotes the eigenvalue of

$$\left. A^{(k)}_{i+1/2} = A(w_i + s_k(w_{i+1} - w_i))\right)$$

with maximum modulus.

Clearly, a kind of AVM methods can be obtained as a particular case of approximate AVM-DOT solvers, simply by taken $q = 1$ and $\omega_1 = 1$.

Jacobian-Free Implementation

We end this section with some notes on the Jacobian-free implementation of AVM-DOT solvers. As it was already mentioned at the end of Sect. 3, this is only possible for polynomial-based methods. To clarify the process, we will focus on the case of an AVM-DOT solver based on internal polynomial approximations.

As it was indicated in [12], the explicit form of $p_n(x)$ combined with Horner's method will be considered instead of the recursive form (6). On the other hand, notice that it will not be necessary to compute the viscosity matrix $Q_{i+1/2}$ explicitely, but only the vector $Q_{i+1/2}(w_{i+1} - w_i)$ appearing in the numerical flux (3).

To ilustrate the procedure, consider the polynomial

$$p_2(x) = \alpha_0 x^4 + \alpha_1 x^2 + \alpha_2 = x^2(\alpha_0 x^2 + \alpha_1) + \alpha_2,$$

where the coefficients are given by $\alpha_0 = -1/8$, $\alpha_1 = 3/4$ and $\alpha_2 = 3/8$. Let $A \equiv A(w)$ be the Jacobian matrix of F evaluated at an intermediate state w, and let v be an arbitrary state; for simplicity, assume that $\lambda_{\max} = 1$. Then the following approximation holds:

$$|A|v \approx p_2(A)v = (A^2(\alpha_0 A^2 + \alpha_1 I) + \alpha_2 I)v.$$

The above expression can be computed using Horner's algorithm:

- Define $v_0 = v$ and compute $\widetilde{v}_0 = A^2 v_0$.
- Calculate $v_1 = \alpha_0 \widetilde{v}_0 + \alpha_1 v_0$ and $\widetilde{v}_1 = A^2 v_1$.
- Compute $v_2 = \widetilde{v}_1 + \alpha_2 v_0$. Then, $|A(w)|v \approx p_2(A)v = v_2$.

The product $A(w)v$ can be approximated using finite differences:

$$A(w)v \approx \frac{F(w + \varepsilon v) - F(w)}{\varepsilon},$$

which leads to

$$A(w)^2 v \approx \frac{F\big(w + F(w + \varepsilon v) - F(w)\big) - F(w)}{\varepsilon} \equiv \Phi_\varepsilon(w; v),$$

where, in practice, the value ε should be chosen small relative to the norm of w. Finally, the vector $|A(w)|v$ can be approximated using the following steps, in which only vector operations and evaluations of the physical flux F are needed:

- Define $v_0 = v$ and compute $\widetilde{v}_0 = \Phi_\varepsilon(w; v_0)$.
- Calculate $v_1 = \alpha_0 \widetilde{v}_0 + \alpha_1 v_0$ and $\widetilde{v}_1 = \Phi_\varepsilon(w; v_1)$.
- Compute $v_2 = \widetilde{v}_1 + \alpha_2 v_0$. Then, $|A(w)|v \approx v_2$.

The detailed procedure to build a Jacobian-free AVM-DOT solver with $p_n(x)$ as basis function is given next. The key point is to approximate the vectors

$$\widetilde{P}^{(k)}_{i+1/2}\Delta w = \left|\lambda^{(k)}_{i+1/2,\max}\right| p_n(A^{(k)})\Delta w, \quad k = 1, \ldots, q,$$

where $\Delta w = w_{i+1} - w_i$ and $A^{(k)} = \left|\lambda^{(k)}_{i+1/2,\max}\right|^{-1} A(w^{(k)}_i)$, with $w^{(k)}_i = w_i + s_k \Delta w$. Assuming that the coefficients α_i of the polynomial $p_n(x)$ have already been computed, the polynomial $p_n(x)$ can be written as

$$p_n(x) = \alpha_0 x^{2(n+1)} + \alpha_1 x^{2n} + \alpha_2 x^{2(n-1)} + \cdots + \alpha_n x^2 + \alpha_{n+1}.$$

Then each term $\widetilde{P}^{(k)}_{i+1/2}\Delta w$ can be approximated using the following algorithm:

- Define $v_0 = \Delta w$ and compute $\widetilde{v}_0 = \left|\lambda^{(k)}_{i+1/2,\max}\right|^{-2} \Phi_\varepsilon(w^{(k)}_i; v_0)$.
- Calculate $v_1 = \alpha_0 \widetilde{v}_0 + \alpha_1 v_0$.
- For $j = 1, \ldots, n$, define $\widetilde{v}_j = \left|\lambda^{(k)}_{i+1/2,\max}\right|^{-2} \Phi_\varepsilon(w^{(k)}_i; v_j)$ and compute $v_{j+1} = \widetilde{v}_j + \alpha_{j+1} v_0$.
- Finally, $\widetilde{P}^{(k)}_{i+1/2}\Delta w \approx \left|\lambda^{(k)}_{i+1/2,\max}\right| v_{n+1}$.

4 The Nonconservative Case

The AVM and AVM-DOT solvers introduced in the previous sections can be extended in a natural way to the case of nonconservative hyperbolic systems. We will focus in this section in AVM-DOT solvers, as they are more general. However, all the results can be readily adapted to AVM solvers.

Consider a hyperbolic system in nonconservative form

$$\partial_t W + \mathcal{A}(W)\partial_x W = 0, \tag{10}$$

where the matrix $\mathcal{A}(W)$ is strictly hyperbolic for each state W belonging to an open convex subset $\Omega \subset \mathbb{R}^M$. The definition of the nonconservative product $\mathcal{A}(W)\partial_x W$ depends on the choice of a family of paths $\Phi(s; W_L, W_R)$ joining arbitrary states W_L and W_R in the phase space Ω: see [22, 28] for details.

The solutions of (10) can be numerically approximated by means of *path-conservative* finite volume schemes of the form [28]:

$$W^{n+1}_i = W^n_i - \frac{\Delta t}{\Delta x}(\mathcal{D}^+_{i-1/2} + \mathcal{D}^-_{i+1/2}), \tag{11}$$

where $\mathcal{D}^\pm_{i+1/2} = \mathcal{D}^\pm(W^n_i, W^n_{i+1})$. Here $\mathcal{D}^-$ and $\mathcal{D}^+$ are two continuous functions from $\Omega \times \Omega$ to Ω satisfying

$$\mathcal{D}^\pm(W, W) = W, \quad \forall\, W \in \Omega,$$

and

$$\mathcal{D}^-(W_0, W_1) + \mathcal{D}^+(W_0, W_1) = \int_0^1 \mathcal{A}(\Phi(s; W_0, W_1)) \frac{\partial \Phi}{\partial s}(s; W_0, W_1)\, ds \quad (12)$$

for every $W_0, W_1 \in \Omega$, with $\Phi(0; W_0, W_1) = W_0$ and $\Phi(1; W_0, W_1) = W_1$. In particular, the *generalized Roe's scheme* [27] is defined by choosing

$$\mathcal{D}^{\pm}_{i+1/2} = \frac{1}{2}\big(\mathcal{A}_\Phi(W_i^n, W_{i+1}^n) \pm |\mathcal{A}_\Phi(W_i^n, W_{i+1}^n)|\big)(W_{i+1}^n - W_i^n),$$

where $\mathcal{A}_\Phi$ is a Roe linearization associated to $\mathcal{A}$ and Φ. In this case, the term $|\mathcal{A}_\Phi(W_i^n, W_{i+1}^n)|$ plays the role of a viscosity matrix. Using Roe's property, it is possible to write the above expression as

$$\mathcal{D}^{\pm}_{i+1/2} = \frac{1}{2}\int_0^1 \mathcal{A}(\Phi(s; W_i^n, W_{i+1}^n)) \frac{\partial \Phi}{\partial s}(s; W_i^n, W_{i+1}^n)\, ds \pm \frac{1}{2}|\mathcal{A}_\Phi(W_i^n, W_{i+1}^n)|(W_{i+1}^n - W_i^n).$$

Then, it is natural to define the Osher-Solomon scheme for solving the nonconservative system (10) as (11) with

$$\mathcal{D}^{\pm}_{i+1/2} = \frac{1}{2}\int_0^1 \mathcal{A}(\Phi(s; W_i^n, W_{i+1}^n)) \frac{\partial \Phi}{\partial s}(s; W_i^n, W_{i+1}^n)\, ds \pm \frac{1}{2}\int_0^1 |\mathcal{A}(\Phi(s; W_i^n, W_{i+1}^n))| \frac{\partial \Phi}{\partial s}(s; W_i^n, W_{i+1}^n)\, ds, \quad (13)$$

or, equivalently,

$$\mathcal{D}^{\pm}_{i+1/2} = \frac{1}{2}\mathcal{A}_\Phi(W_i^n, W_{i+1}^n)(W_{i+1}^n - W_i^n) \pm \frac{1}{2}\int_0^1 |\mathcal{A}(\Phi(s; W_i^n, W_{i+1}^n))| \frac{\partial \Phi}{\partial s}(s; W_i^n, W_{i+1}^n)\, ds. \quad (14)$$

Notice that (13) is more general than (14), as the latter relies on the existence of a Roe linearization $\mathcal{A}_\Phi$. Therefore, (13) could be used in the cases in which a Roe linearization is not known or is difficult to compute.

A good choice of the family of paths Φ may be difficult or very costly in practice, and usually relies on the physics of the problem (see [28]). A simple choice, commonly used in the literature, is given by the family of segments: $\Phi(s; W_L, W_R) = W_L + s(W_R - W_L)$; we will consider this choice throughout the rest of the section. Then, denoting $\mathcal{A}_{i+1/2} = \mathcal{A}_\Phi(W_i^n, W_{i+1}^n)$, we have:

$$\mathcal{D}^{\pm}_{i+1/2} = \frac{1}{2}\left(\mathcal{A}_{i+1/2} \pm \int_0^1 |\mathcal{A}(W_i^n + s(W_{i+1}^n - W_i^n))|\, ds\right)(W_{i+1}^n - W_i^n),$$

where the integral $\int_0^1 |\mathcal{A}(W_i^n + s(W_{i+1}^n - W_i^n))|\, ds$ can be interpreted as a viscosity term. Next, this integral can be approximated using a Gauss-Legendre quadrature formula, which leads to

$$\mathcal{D}^{\pm}_{i+1/2} = \frac{1}{2}\left(\mathcal{A}_{i+1/2} \pm \sum_{k=1}^{q} \omega_k |\mathcal{A}^{(k)}_{i+1/2}|\right)(W_{i+1}^n - W_i^n), \tag{15}$$

where $\mathcal{A}^{(k)}_{i+1/2} = \mathcal{A}(W_i^n + s_k(W_{i+1}^n - W_i^n))$. Therefore, (15) can be interpreted as a nonconservative extension of the DOT numerical flux. This approach has also been considered in [16].

Once formula (15) has been derived, AVM-DOT schemes for the nonconservative system (10) can be built in a natural way, considering

$$\mathcal{D}^{\pm}_{i+1/2} = \frac{1}{2}\left(\mathcal{A}_{i+1/2} \pm \sum_{k=1}^{q} \omega_k \widetilde{P}^{(k)}_{i+1/2}\right)(W_{i+1}^n - W_i^n),$$

where $\widetilde{P}^{(k)}_{i+1/2}$ is defined in (9).

We will focus now in the particular case of a hyperbolic system of conservation laws with source terms and nonconservative products, that is,

$$\partial_t w + \partial_x F(w) + B(w)\partial_x w = G(w)\partial_x H, \tag{16}$$

where $w(x,t) \in \mathcal{O}$ (being $\mathcal{O} \subset \mathbb{R}^N$ open and convex), $F\colon \mathcal{O} \to \mathbb{R}^N$ is a smooth flux function, $B\colon \mathcal{O} \to \mathcal{M}_N(\mathbb{R})$ is a smooth matricial function, and $G\colon \mathcal{O} \to \mathbb{R}^N$ and $H\colon \mathbb{R} \to \mathbb{R}$ are given functions. System (16) can be written in the form (10) adding the trivial equation $\partial_t H = 0$ and defining

$$W = \begin{pmatrix} w \\ H \end{pmatrix} \in \Omega = \mathcal{O} \times \mathbb{R} \subset \mathbb{R}^{N+1}, \quad \mathcal{A}(W) = \begin{pmatrix} A(w) & -G(w) \\ 0 & 0 \end{pmatrix},$$

where $A(w) = \frac{\partial F}{\partial w}(w) + B(w)$. In this case, a Roe linearization $\mathcal{A}_{i+1/2}$ can be defined as [27]

$$\mathcal{A}_{i+1/2} = \begin{pmatrix} A_{i+1/2} & -G_{i+1/2} \\ 0 & 0 \end{pmatrix},$$

where $A_{i+1/2} = \mathcal{L}_{i+1/2} + B_{i+1/2}$, $\mathcal{L}_{i+1/2}$ being a Roe matrix for the flux F in the usual sense, that is, $\mathcal{L}_{i+1/2}(w_{i+1}^n - w_i^n) = F(w_{i+1}^n) - F(w_i^n)$; $B_{i+1/2}$ is a matrix verifying

$$B_{i+1/2}(w_{i+1}^n - w_i^n) = \left(\int_0^1 B(w_i^n + s(w_{i+1}^n - w_i^n))\, ds\right)(w_{i+1}^n - w_i^n),$$

and $G_{i+1/2}$ is a vector satisfying

$$G_{i+1/2}(H_{i+1} - H_i) = \left(\int_0^1 G(w_i^n + s(w_{i+1}^n - w_i^n))\, ds \right)(H_{i+1} - H_i).$$

A simple calculation gives

$$|\mathcal{A}(W)| = \begin{pmatrix} |A(w)| & -|A(w)|A(w)^{-1}G(w) \\ 0 & 0 \end{pmatrix},$$

as long as $A(w)$ is nonsingular. Substituing in (15), the DOT scheme for solving (16) can be written as

$$w_{i+1}^n = w_i^n - \frac{\Delta t}{\Delta x}(D_{i-1/2}^+ + D_{i+1/2}^-), \tag{17}$$

with

$$D_{i+1/2}^{\pm} = \frac{1}{2}\Bigg(F(w_{i+1}^n) - F(w_i^n) + B_{i+1/2}(w_{i+1}^n - w_i^n) - G_{i+1/2}(H_{i+1} - H_i) \\ \pm \sum_{k=1}^{q} \omega_k |A_{i+1/2}^{(k)}| \big(w_{i+1}^n - w_i^n - (A_{i+1/2}^{(k)})^{-1} G_{i+1/2}^{(k)}(H_{i+1} - H_i) \big) \Bigg), \tag{18}$$

where $A_{i+1/2}^{(k)} = A(w_i^n + s_k(w_{i+1}^n - w_i^n))$, and similarly for $G_{i+1/2}^{(k)}$.

Finally, AVM-DOT schemes for (16) are obtained by substituting $|A_{i+1/2}^{(k)}|$ by $\widetilde{P}_{i+1/2}^{(k)}$ in (18):

$$D_{i+1/2}^{\pm} = \frac{1}{2}\Bigg(F(w_{i+1}^n) - F(w_i^n) + B_{i+1/2}(w_{i+1}^n - w_i^n) - G_{i+1/2}(H_{i+1} - H_i) \\ \pm \sum_{k=1}^{q} \omega_k \widetilde{P}_{i+1/2}^{(k)} \big(w_{i+1}^n - w_i^n - (A_{i+1/2}^{(k)})^{-1} G_{i+1/2}^{(k)}(H_{i+1} - H_i) \big) \Bigg). \tag{19}$$

In the case of a system of conservation laws (that is, $B = 0$ and $G = 0$), the scheme (17) can be written in the form (2) by simply taking $F_{i+1/2} = D_{i+1/2}^- + F(w_i^n)$ or, equivalently, $F_{i+1/2} = -D_{i+1/2}^+ + F(w_{i+1}^n)$.

5 Numerical Experiments

In this section we test the performances of AVM-DOT schemes with some challenging problems related to ideal MHD equations in the conservative case, and to multilayer shallow water equations in the nonconservative case.

Depending on the basis function $f(x)$, the AVM-DOT schemes will be denoted:

- DOT-Cheb-$2p$: $f(x)$ is the Chebyshev polynomial $\tau_{2p}(x)$.
- DOT-Newman-r: $f(x)$ is taken as the Newman rational function $R_r(x)$.
- DOT-Halley-r: $f(x) = H_r(x)$, the r-th Halley rational function.
- DOT: the original DOT method in [15], in which the eigendecomposition is computed numerically.

For higher order schemes, third-order PHM [21] reconstructions have been used, combined with a third-order TVD Runge-Kutta method for time stepping.

5.1 Applications to Magnetohydrodynamics

The MHD system of equations is given by [4]

$$\begin{cases} \partial_t \rho = -\nabla \cdot (\rho \mathbf{v}), \\ \partial_t (\rho \mathbf{v}) = -\nabla \cdot \left(\rho \mathbf{v}\mathbf{v}^T + \left(P + \frac{1}{2}\mathbf{B}^2 \right) I - \mathbf{B}\mathbf{B}^T \right), \\ \partial_t \mathbf{B} = \nabla \times (\mathbf{v} \times \mathbf{B}), \\ \partial_t E = -\nabla \cdot \left(\left(\frac{\gamma}{\gamma - 1} P + \frac{1}{2}\rho q^2 \right) \mathbf{v} - (\mathbf{v} \times \mathbf{B}) \times \mathbf{B} \right), \end{cases} \tag{20}$$

where ρ represents the mass density, $\mathbf{v} = (v_x, v_y, v_z)^t$ and $\mathbf{B} = (B_x, B_y, B_z)^t$ are the velocity and magnetic fields, and E is the total energy. Denoting by q and B the magnitudes of the velocity and magnetic fields, the total energy is

$$E = \frac{1}{2}\rho q^2 + \frac{1}{2}B^2 + \rho\varepsilon,$$

where the specific internal energy ε and the hydrostatic pressure P are related through the equation of state $P = (\gamma - 1)\rho\varepsilon$, with γ the adiabatic constant. In addition to the equations, the magnetic field must satisfy the divergence-free condition

$$\nabla \cdot \mathbf{B} = 0. \tag{21}$$

In the numerical experiments, condition (21) has been imposed by means of the projection method in [3]. Notice that for $\mathbf{B} = \mathbf{0}$, system (20) reduces to the Euler equations for gases. The eigenstructure of system (20) is completely determined: see, e.g., [4].

5.1.1 Stationary Contact Discontinuity

The purpose of this test, first proposed in [17], is to study the effect of the numerical diffusion in the approximation of a stationary contact discontinuity. This effect, known as *numerical heat conduction*, may cause incorrect heating across the discontinuity. The initial conditions for the Euler equations are given by

$$(\rho, v_x, P) = \begin{cases} (1, 0, 1) & \text{for } x \leq 0.5, \\ (2, 0, 1) & \text{for } x > 0.5, \end{cases}$$

with $\gamma = 1.4$. The solution consists in a stationary contact wave located at $x = 0.5$. The problem has been solved in the domain $[0, 1]$ with 200 cells and CFL$= 0.5$ until a final time $t = 4$.

Figure 3 shows the approximations to the density component. Both in the first- and third-order solutions, DOT-Newman-4 gives the best approximation to the solution, followed by DOT-Halley-2, DOT-Cheb-4 and DOT-Halley-1. The HLL scheme gives a very diffusive resolution of the discontinuity. On the other hand, Fig. 4 shows the corresponding efficiency curves which represent, in logarithmic scale, the CPU times versus the L^1 errors with respect to the exact solution for different meshes. In any case, DOT-Newman-4 is the most efficient solver. This test shows that the choice of

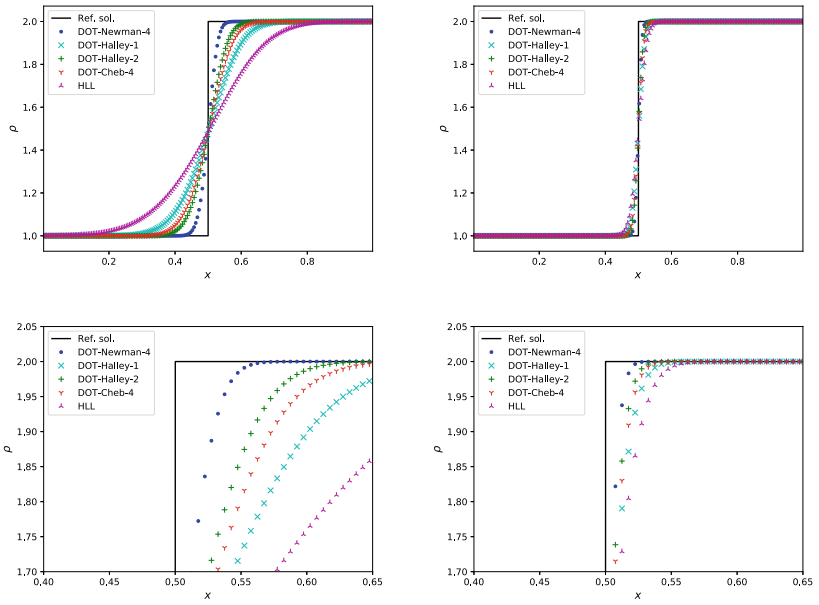


Fig. 3 Test 5.1.1: Left: first order. Right: third order. The solutions obtained with the Roe and DOT schemes coincide with the reference (exact) solution. The lower row shows a zoom near the upper part of the discontinuity

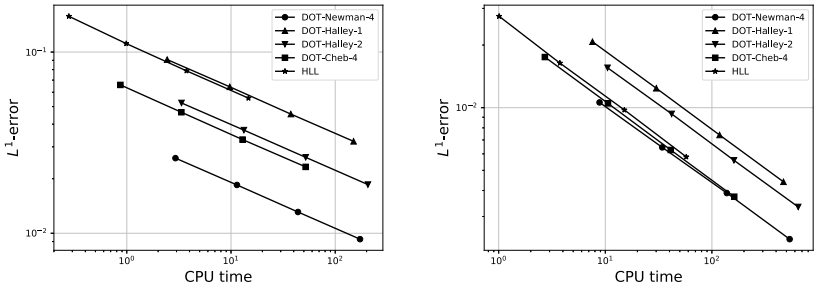


Fig. 4 Test 5.1.1: Efficiency curves CPU vs. L^1-error. Left: first order. Right: third order

an appropriate first-order solver is important even when it is going to be used as a building block for higher-order schemes.

5.1.2 Brio-Wu Shock Tube Problem

This experiment was proposed in [4] to show the formation of a compound wave consisting of a shock followed by a rarefaction wave. The initial conditions are the following:

$$(\rho, v_x, v_y, v_z, B_x, B_y, B_z, P) = \begin{cases} (1, 0, 0, 0, 0.75, 1, 0, 1) & \text{for } x \le 0, \\ (0.125, 0, 0, 0, 0.75, -1, 0, 0.1) & \text{for } x > 0, \end{cases}$$

with $\gamma = 2$. The problem has been solved until time $t = 0.2$ in the interval $[-1, 1]$ with a 1000 cell spatial discretization and CFL$= 0.8$. The results are shown in Fig. 5: in this case there are no appreciable differences between the solutions computed with Roe, DOT-Newman-4, DOT-Halley-2, DOT-Cheb-4 and DOT. On the other hand, the first order HLL method provides a worse resolution of the compound wave, which is however improved in third order.

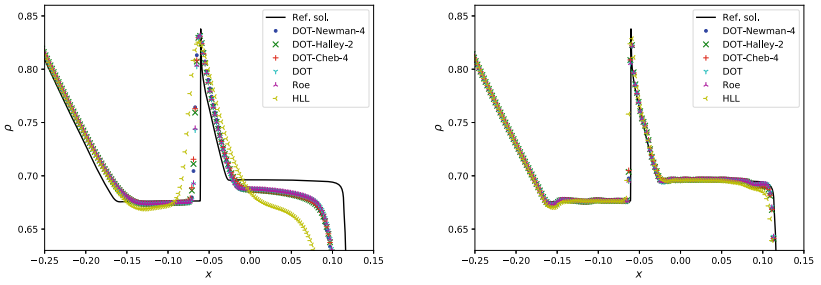


Fig. 5 Test 5.1.2: Zoom of the density compound wave. Left: first order. Right: third order

5.1.3 Orszag-Tang Vortex

The Orszag-Tang vortex [25] constitutes a model of transition to supersonic MHD turbulence in which, departing from a smooth state, complex interactions between shock waves are generated as the system evolves.

For $(x, y) \in [0, 2\pi] \times [0, 2\pi]$, the initial data are given by

$$\rho(x, y, 0) = \gamma^2, \quad v_x(x, y, 0) = -\sin(y), \quad v_y(x, y, 0) = \sin(x),$$
$$B_x(x, y, 0) = -\sin(y), \quad B_y(x, y, 0) = \sin(2x), \quad P(x, y, 0) = \gamma,$$

with $\gamma = 5/3$. Periodic boundary conditions are imposed in the x- and y-directions. The computations have been done using a 192×192 uniform mesh and CFL = 0.8.

Figure 6 shows the results obtained with the third-order DOT-Cheb-4 scheme at time $t = 3$; similar solutions are obtained with the third-order DOT-Newman-4, DOT-Halley-2, and DOT schemes. The results are in very good agreement with those found in the literature, thus showing that our schemes are robust and accurate enough to resolve the complicated structure of this vortex system. Finally, Table 1 shows the relative CPU times with respect to the first-order DOT scheme.

The relativistic version of the Orszag-Tang problem has also been considered. In this case, due to the very complex form of the Jacobians of the system, the Jacobian-free implementation introduced at the end of Sect. 3 has been found to be a very advantageous choice. Figure 7 shows the solution computed at time $t = 4$ with a

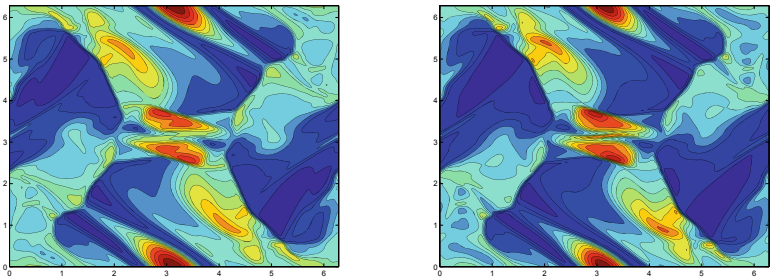


Fig. 6 Test 5.1.3: Density (left) and pressure (right) computed at time $t = 3$ with the third-order DOT-Cheb-4 scheme

Table 1 Test 5.1.3: Relative CPU times with respect to the first-order DOT solver

Method	CPU (first order)	CPU (third order)
DOT	1.00	5.82
DT-Cheb-4	0.16	1.04
DOT-Newman-4	0.38	2.32
DOT-Halley-2	0.50	2.79
HLL	0.05	0.36

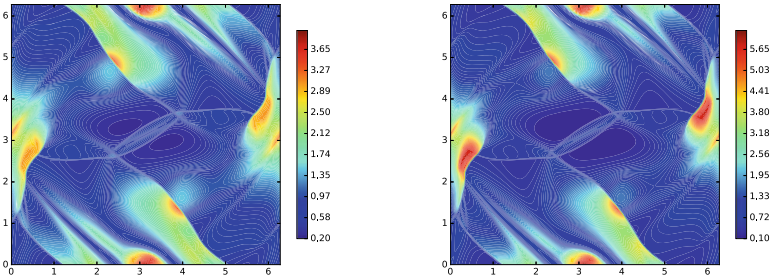


Fig. 7 Relativistic Orszag-Tang vortex: Left: density. Right: pressure

Jacobian-free second-order PVM-int-8 scheme, based on the internal polynomial approximation introduced in Sect. 2. At a qualitative level, our results are in good agreement with those found in, e.g., [32]. For a more detailed discussion about Jacobian-free AVM-DOT solvers applied to relativistic MHD, the reader is referred to [12].

5.1.4 The Rotor Problem

In this section we consider the rotor problem proposed in [1]. At the beginning, a dense disk rotates at the center of the domain, while the ambient fluid remains at rest. These two areas are connected with a taper function, which helps to reduce the initial discontinuity. As time evolves, the rotating dense fluid tends to be confined into an oblate shape, due to the action of the magnetic field.

The computational domain is $[0, 1] \times [0, 1]$ with periodic boundary conditions. Defining $r_0 = 0.1$, $r_1 = 0.115$, $f = (r_1 - r)/(r_1 - r_0)$ and $r = [(x - 0.5)^2 + (y - 0.5)^2]^{1/2}$, the initial conditions are given by

$$(\rho(x, y), v_x(x, y), v_y(x, y)) = \begin{cases} (10, -(y - 0.5)/r_0, (x - 0.5)/r_0) & \text{if } r < r_0, \\ (1 + 9f, -(y - 0.5)f/r, (x - 0.5)f/r) & \text{if } r_0 < r < r_1, \\ (1, 0, 0) & \text{if } r > r_1, \end{cases}$$

with $B_x = 2.5/\sqrt{4\pi}$, $B_y = 0$ and $P = 0.5$. We take $\gamma = 5/3$.

Figure 8 shows the solutions obtained with the third-order DOT-Cheb-4 scheme at time $t = 0.295$ on a 200×200 mesh with CFL$= 0.8$. The results are in good agreement with those presented in [1]. As in the previous test, DOT-Newman-4 and DOT-Halley-2 give similar results as DOT-Cheb-4. On the contrary, the DOT scheme fails for this problem around time $t \approx 0.187$. Finally, the third-order HLL and DOT-Cheb-4 methods are compared in Fig. 9. As it can be seen, DOT-Cheb-4 produces more precise results than HLL, which indicates that the choice of a precise first-order solver is important even when designing high-order schemes.

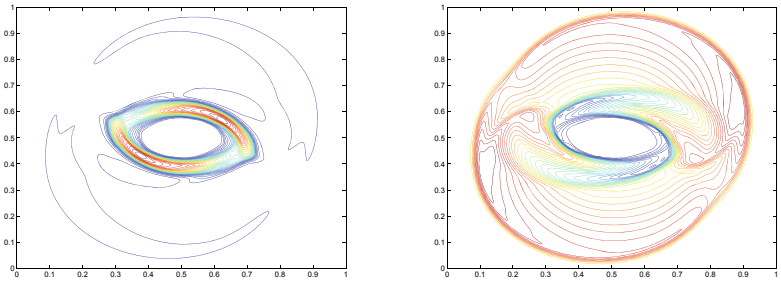


Fig. 8 Test 5.1.4: Density ρ (left) and pressure P (right) computed at time $t = 0.295$ with the third-order DOT-Cheb-4

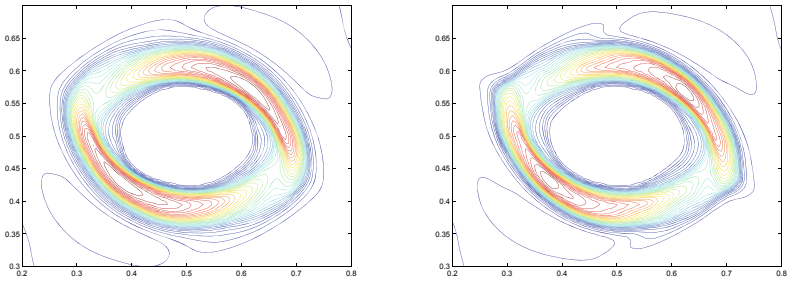


Fig. 9 Test 5.1.4: Comparison between the density solutions obtained with the third-order HLL (left) and DOT-Cheb-4 (right) schemes

5.2 Applications to the Two-Layer Shallow Water System

The two-layer shallow water equations constitute a representative model of the non-conservative systems considered in Sect. 4, as they include both source and non-conservative coupling terms (see [7]). The equations governing the one-dimensional flow of two superposed inmiscible layers of shallow water fluids can be written in the form (16) by taking

$$w = \begin{pmatrix} h_1 \\ q_1 \\ h_2 \\ q_2 \end{pmatrix}, \quad F(w) = \begin{pmatrix} q_1 \\ \dfrac{q_1^2}{h_1} + \dfrac{g}{2}h_1^2 \\ q_2 \\ \dfrac{q_2^2}{h_2} + \dfrac{g}{2}h_2^2 \end{pmatrix}, \quad G(w) = \begin{pmatrix} 0 \\ gh_1 \\ 0 \\ gh_2 \end{pmatrix}, \quad B(w) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & gh_1 & 0 \\ 0 & 0 & 0 & 0 \\ rgh_2 & 0 & 0 & 0 \end{pmatrix},$$

where h_j are the fluid depths, $q_j = h_j u_j$ represent the discharges, u_j are the velocities, and $H(x)$ is the depth function measured from a fixed level of reference; g is the gravity constant and $r = \rho_1/\rho_2$ is the ratio of densities. Notice that $j = 1$ corresponds to the upper layer and $j = 2$ to the lower one.

To build the AVM-DOT fluxes (19), we define

$$B_{i+1/2} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & gh_{1,i+1/2} & 0 \\ 0 & 0 & 0 & 0 \\ rgh_{2,i+1/2} & 0 & 0 & 0 \end{pmatrix}, \quad G_{i+1/2} = \begin{pmatrix} 0 \\ gh_{1,i+1/2} \\ 0 \\ gh_{2,i+1/2} \end{pmatrix},$$

where

$$h_{k,i+1/2} = \frac{h_{k,i} + h_{k,i+1}}{2}, \quad k = 1, 2.$$

Notice that the exact eigenstructure of the two-layer system is not explicitely known. However, a first order approximation of the maximum wave speed is given by

$$|\lambda_{i+1/2,\max}| \approx |\bar{u}_{i+1/2}| + c_{i+1/2},$$

where

$$\bar{u}_{i+1/2} = \frac{q_{1,i+1/2} + q_{2,i+1/2}}{h_{1,i+1/2} + h_{2,i+1/2}}, \quad c_{i+1/2} = \sqrt{g(h_{1,i+1/2} + h_{2,i+1/2})}.$$

5.2.1 Internal Dam-Break

This test was proposed in [9] to simulate a dam-break in a two-layer system. The initial conditions are given by

$$h_1(x, 0) = \begin{cases} 0.9 & \text{if } x < 5, \\ 0.1 & \text{if } x \geq 5, \end{cases} \qquad h_2(x, 0) = 1 - h_1(x, 0),$$

and $q_1(x, 0) = q_2(x, 0) = 0$, for $x \in [0, 10]$. The ratio of densities is taken as $r = 0.99$. The problem has been solved using a mesh with 200 grid points until time $t = 20$, with CFL number 0.9. Open boundary conditions have been imposed.

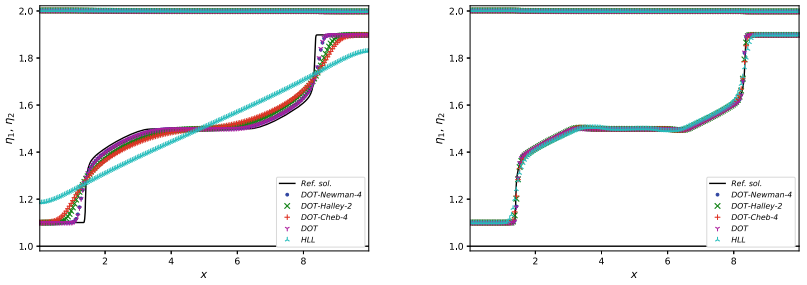


Fig. 10 Test 5.2.1: Free surface and interface. Left: first order. Right: third order

Figure 10 shows the free surface and the interface ($\eta_j = h_j - H$). The best results in first order are obtained with the DOT-Newman-4 and DOT schemes, followed by DOT-Halley-2 and DOT-Cheb-4, while HLL is not able to capture the interface correctly. On the other hand, in third order all the schemes perform equally well, being HLL the one that gives the less precise results.

5.2.2 Transcritical Flux with Shock

The initial condition for this test consists in an internal dam-break over a non-flat bottom, which eventually tends towards a stationary transcritical solution with a shock (see [10]). Specifically, the initial conditions are given by $q_1(x, 0) = q_2(x, 0) = 0$,

$$h_1(x, 0) = \begin{cases} 0.48 & \text{for } x < 0, \\ 0.02 & \text{for } x \geq 0, \end{cases} \qquad h_2(x, 0) = H(x) - h_1(x, 0),$$

and the bottom topography is defined by

$$H(x) = 1 - \frac{1}{2} e^{-x^2}, \quad x \in [-5, 5].$$

Open wall boundary conditions have been imposed, and the ratio of densities has been chosen as $r = 0.998$.

The numerical solutions have been computed on a mesh with 200 grid points until final time $t = 100$, with CFL number 0.9. The results have been represented in Fig. 11. In first order, the DOT-Newman-4 and DOT schemes provide the best resolution of the interface, followed by DOT-Halley-2 and DOT-Cheb-4; on the other hand, HLL is unable to resolve the complex structure of the interface. The situation improves when going to third order, although again HLL presents a worse resolution near discontinuities. This can be better seen in the bottom row of Fig. 11, where a closer view of the shock has been plotted. Notice also that the DOT scheme presents more pronounced oscillations near the shock than DOT-Newman-4. Finally, the relative CPU times with respect to the first-order DOT scheme are shown in Table 2.

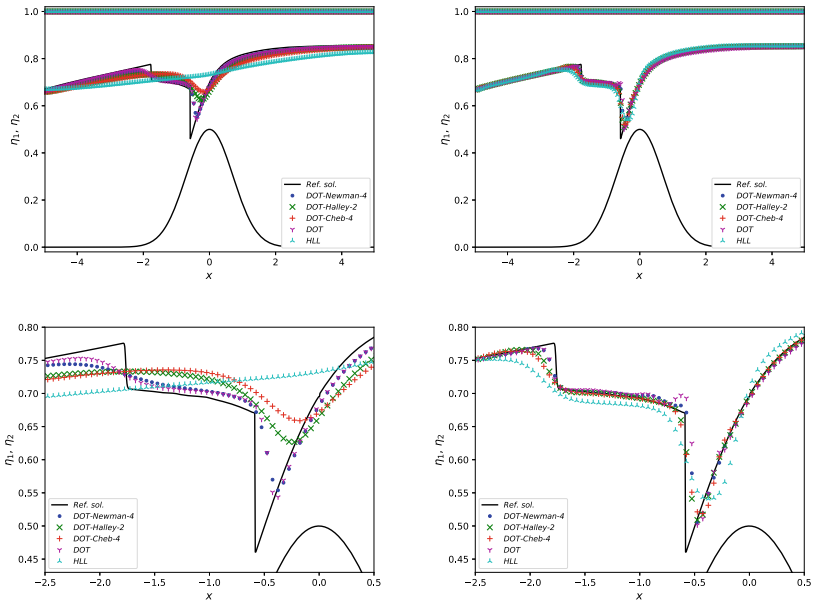


Fig. 11 Test 5.2.2: Free surface, interface and bottom. Left: first order. Right: third order. The lower row shows a closer view of the shock at the interface

Table 2 Test 5.2.2: Relative CPU times with respect to the first-order DOT solver

Method	CPU (first order)	CPU (third order)
DOT	1.00	2.98
DOT-Cheb-4	0.15	0.51
DOT-Newman-4	0.37	1.17
DOT-Halley-2	0.44	1.39
HLL	0.06	0.24

Acknowledgements This research has been partially supported by the Spanish Government Research project RTI2018-096064-B-C21. The numerical computations have been performed at the Laboratory of Numerical Methods of the University of Málaga.

References

1. Balsara, D.S., Spicer, D.S.: A staggered mesh algorithm using high order Godunov fluxes to ensure solenoidal magnetic fields in magnetohydrodynamic simulations. J. Comput. Phys. **149**, 270–292 (1999)
2. Bernstein, S.: Sur la meilleure approximation de $|x|$ par del polynômes de degrés donés. Acta Math. **37**, 1–57 (1913)

3. Brackbill, J.U., Barnes, J.C.: The effect of nonzero $\nabla \cdot B$ on the numerical solution of the magnetohydrodynamic equations. J. Comput. Phys. **35**, 426–430 (1980)
4. Brio, M., Wu, C.C.: An upwind differencing scheme for the equations of ideal magnetohydrodynamics. J. Comput. Phys. **75**, 400–422 (1988)
5. Candela, V., Marquina, A.: Recurrence relations for rational cubic methods I: the Halley method. Computing **44**, 169–184 (1990)
6. Cargo, P., Gallice, G.: Roe matrices for ideal MHD and systematic construction of Roe matrices for systems of conservation laws. J. Comput. Phys. **136**, 446–466 (1997)
7. Castro, M.J., Macías, J., Parés, C.: A Q-scheme for a class of systems of coupled conservation laws with source term. Application to a two-layer 1-D shallow water system. Math. Mod. Num. Anal. **35**, 107–127 (2001)
8. Castro Díaz, M.J., Fernández-Nieto, E.D.: A class of computationally fast first order finite volume solvers: PVM methods. SIAM J. Sci. Comput. **34**, A2173–A2196 (2012)
9. Castro Díaz, M.J., Fernández-Nieto, E.D., Narbona-Reina, G., de la Asunción, M.: A second order PVM flux limiter method. Application to magnetohydrodynamics and shallow stratified flows. J. Comput. Phys. **262**, 172–193 (2014)
10. Castro, M.J., Gallardo, J.M., Marquina, A.: A class of incomplete Riemann solvers based on uniform rational approximations to the absolute value function. J. Sci. Comput. **60**, 363–389 (2014)
11. Castro, M.J., Gallardo, J.M., Marquina, A.: Approximate Osher-Solomon schemes for hyperbolic systems. Appl. Math. Comput. **272**, 347–368 (2016)
12. Castro, M.J., Gallardo, J.M., Marquina, A.: Jacobian-free approximate solvers for hyperbolic systems: Application to relativistic magnetohydrodynamics. Comput. Phys. Commun. **219**, 108–120 (2017)
13. Cordier, F., Degond, P., Kumbaro, A.: Phase appearance or disappearance in two-phase flows. J. Sci. Comput. **58**, 115–148 (2013)
14. Degond, P., Peyrard, P.F., Russo, G., Villedieu, P.: Polynomial upwind schemes for hyperbolic systems. C. R. Acad. Sci. Paris Sér. **I**(328), 479–483 (1999)
15. Dumbser, M., Toro, E.F.: On universal Osher-type schemes for general nonlinear hyperbolic conservation laws. Commun. Comput. Phys. **10**, 635–671 (2011)
16. Dumbser, M., Toro, E.F.: A simple extension of the Osher Riemann solver to non-conservative hyperbolic systems. J. Sci. Comput. **48**, 70–88 (2011)
17. Einfeldt, B., Munz, C.D., Roe, P.L., Sjögreen, B.: On Godunov-type methods near low densities. J. Comput. Phys. **92**, 273–295 (1991)
18. Gallardo, J.M., Schneider, K.A., Castro, M.J.: On a class of two-dimensional incomplete Riemann solvers. J. Comput. Phys. **386**, 541–567 (2019)
19. Godunov, S.K.: Finite difference methods for the computation of discontinuous solutions of the equations of fluid dynamics. Math. USSR Sbornik **47**, 271–306 (1959)
20. Harten, A., Lax, P.D., van Leer, B.: On upstream differencing and Godunov type schemes for hyperbolic conservation laws. SIAM Rev. **25**, 35–61 (1983)
21. Marquina, A.: Local piecewise hyperbolic reconstructions for nonlinear scalar conservation laws. SIAM J. Sci. Comput. **15**, 892–915 (1994)
22. dal Maso, G., LeFloch, P.G., Murat, F.: Definition and weak stability of nonconservative products. J. Math. Pures Appl. **74**, 483–548 (1995)
23. Ndjinga, M., Kumbaro, A., de Vuyst, F., Laurent-Gengoux, P.: Numerical simulation of hyperbolic two-phase flow models using a Roe-type solver. Nucl. Eng. Des. **238**, 2075–2083 (2008)
24. Newman, D.J.: Rational approximation to $|x|$. Michigan Math. J. **11**, 11–14 (1964)
25. Orszag, S.A., Tang, C.M.: Small scale structure of two-dimensional magnetohydrodynamic turbulence. J. Fluid Mech. **90**, 129–143 (1979)
26. Osher, S., Solomon, F.: Upwind difference schemes for hyperbolic conservation laws. Math. Comput. **38**, 339–374 (1982)
27. Parés, C., Castro, M.J.: On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems. M2AN **38**, 821–852 (2004)

28. Parés, C.: Numerical methods for nonconservative hyperbolic systems: a theoretical framework. SIAM J. Num. Anal. **44**, 300–321 (2006)
29. Roe, P.L.: Approximate Riemann solvers, parameter vectors, and difference schemes. J. Comput. Phys. **43**, 357–372 (1981)
30. Toro, E.F.: Riemann Solvers and Numerical Methods for Fluid Dynamics. Springer, Heidelberg (2009)
31. Torrilhon, M.: Krylov-Riemann solver for large hyperbolic systems of conservation laws. SIAM J. Sci. Comput. **34**, A2072–A2091 (2012)
32. Zanotti, O., Dumbser, M.: A high order special relativistic hydrodynamic and magnetohydrodynamic code with space-time adaptive mesh refinement. Comput. Phys. Commun. **188**, 110–127 (2015)

Entropy–Based Methods for Uncertainty Quantification of Hyperbolic Conservation Laws

Martin Frank, Jonas Kusch, and Jannick Wolters

Abstract Using standard intrusive techniques when solving hyperbolic conservation laws with uncertainties can lead to oscillatory solutions as well as non-hyperbolic moment systems. Entropy-based Stochastic Galerkin methods, on the other hand, guarantee hyperbolicity and entropy decay. A key challenge facing these methods is computational cost, since they require repeatedly solving a non-linear optimization problem. Furthermore, the spatial and temporal discretization needs to preserve realizability, meaning that the existence of a unique solution to the optimization problem must be ensured. We review strategies to guarantee realizability, which use a special choice of the numerical flux while considering errors from the optimization solve. Most importantly, we indicate how intrusive entropy-based closures can be made competitive. We show several numerical test cases and discuss the advantages and disadvantages of several uncertainty propagation methods.

1 Introduction

Hyperbolic equations play an important role in various research as well as industrial areas. The most common equations of this kind model the behavior of liquids, gases and plasmas and are thus widely used in the automotive and aerospace industry. Because of this popularity, many highly efficient and robust implementations for these models are available. The respective codes have shown to, for example,

M. Frank (✉) · J. Kusch · J. Wolters
Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
e-mail: martin.frank@kit.edu

J. Kusch
e-mail: jonas.kusch@kit.edu

J. Wolters
e-mail: jannick.wolters@kit.edu

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*, SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2_2

simulate the airflow around an airfoil very precisely, but only if the provided input data is identical or at least extremely close to the experimental setup. Any arising uncertainties in the input parameters, originating from e.g. measurement tolerances, imperfect information or modeling assumptions cannot be represented and thus lead to differences in the results of experiments and simulations. Therefore, propagating these uncertainties through complex partial differential equations has become an important topic in the last decades.

We consider a parameterized system of hyperbolic equations of the form

$$\partial_t \boldsymbol{u}(t, \boldsymbol{x}, \boldsymbol{\xi}) + \nabla \cdot \boldsymbol{f}(\boldsymbol{u}(t, \boldsymbol{x}, \boldsymbol{\xi})) = \boldsymbol{0} \quad \text{in } D\,, \tag{1a}$$

$$\boldsymbol{u}(t = 0, \boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{u}_{\mathrm{IC}}(\boldsymbol{x}, \boldsymbol{\xi}) \tag{1b}$$

with state variable $\boldsymbol{u} \in \mathcal{D} \subset \mathbb{R}^m$ depending on time $t \in \mathbb{R}^+$, spatial position $\boldsymbol{x} \in D \subset \mathbb{R}^d$ and uncertain parameter $\boldsymbol{\xi} \in \Theta \subseteq \mathbb{R}^p$. The physical flux is given by $\boldsymbol{f} : \mathcal{D} \to \mathbb{R}^{d \times m}$. Note, that for ease of notation, the uncertainties $\boldsymbol{\xi}$ here only enter through the initial condition, i.e. only the initial data is subject to randomness. Boundary conditions are omitted for now as they are specific to the studied problem and will be supplied for the individual test cases in the later sections. We assume that all random parameters are independent with a joint probability density function $f_\Xi = \prod_{i=1}^{p} f_{\Xi,i}(\xi_i)$.

As the solution of (1a) is now subject to randomness, one is often interested in determining the statistical moments of the solution, where the first and second order moments, i.e. the mean and variance of $\boldsymbol{u}$, given by

$$\mathrm{E}[\boldsymbol{u}] = \langle \boldsymbol{u} \rangle, \qquad \mathrm{Var}[\boldsymbol{u}] = \langle (\boldsymbol{u} - \mathrm{E}[\boldsymbol{u}])^2 \rangle$$

are usually most interesting. We define the bracket operator above as

$$\langle \cdot \rangle := \int_\Theta \cdot f_\Xi(\boldsymbol{\xi})\, d\xi_1 \dots d\xi_p.$$

Generally, the methods of Uncertainty Quantification (UQ) can be divided into two groups, intrusive and non-intrusive, meaning the methods either require an intrusive change of an existing deterministic solver, or the existing code can be repurposed in a black-box manner. Several textbooks on UQ have appeared in recent years [5, 10, 27, 28, 35, 40], and we refer the reader to these textbooks for a general overview and references to the original works. In this paper, we nevertheless want to shed some light on how these methods can be expected to perform when applied to hyperbolic conservation laws. We will also make some general statements that we believe are not well-known in some part of the literature. As it has been shown [33], the stochastic Galerkin method cannot be applied directly to conservation laws because it is prone to yield oscillatory solutions that might result in the loss of hyperbolicity and e.g. negative densities. We then discuss in detail the Intrusive Polynomial Moment (IPM) method [33], which can be seen as a generalization to the stochastic Galerkin method.

This method is based on entropy minimization, and choosing a suitable entropy guarantees hyperbolicity. On the other hand, the method comes at the cost of solving an optimization problem at every point in time for every spatial cell.

The IPM method has been inspired by entropy-based closures in kinetic theory. In fact, the moment closure problem for a kinetic equation in velocity space is very similar to treating uncertain parameters [20]. In the case of kinetic equations, the techniques to put entropy-based closures into practice have been refined in a series of papers [1–3, 9, 14, 29], and the application to UQ draws from this experience. In this paper, we discuss realizability-preserving spatial discretizations, and acceleration techniques to solve the IPM system more efficiently. This review is based on the papers [19, 20, 23].

2 Why Galerkin-Type Intrusive Methods?

For the following discussion, we assume that we want to approximate the expected value

$$\mathrm{E}[\boldsymbol{u}] = \langle \boldsymbol{u} \rangle = \int_{\Theta} \boldsymbol{u} f_{\Xi} \, d\xi_1 \dots d\xi_p$$

of the solution $\boldsymbol{u}$ at a given time t as a function of $\boldsymbol{x}$. Since the expected value is an integral of the solution against the probability density function, all non-intrusive UQ methods can be understood and analyzed as numerical quadrature rules

$$\mathrm{E}[\boldsymbol{u}](t,x) \approx \sum_{k=1}^{N} w^k \boldsymbol{u}(t, \boldsymbol{x}, \boldsymbol{\xi^k}) \ .$$

- Monte-Carlo (MC) methods sample $\boldsymbol{\xi^k}$ from the probability density function f_{Ξ} and take $w^k = \frac{1}{N}$.
- Number-theoretic/Quasi-Monte Carlo (QMC) methods for uniformly distributed random variables use a low-discrepancy sequence for $\boldsymbol{\xi^k}$ and again $w^k = \frac{1}{N}$.
- Tensorized quadrature rules take one-dimensional (e.g. Gaussian) quadrature rules for each random input ξ_i. The grid for the $\boldsymbol{\xi^k}$ is defined as a Cartesian product of the one-dimensional grids and the weights are products of the one-dimensional weights.
- Sparse grid quadrature rules use nodes $\boldsymbol{\xi^k}$ on a (e.g. Smolyak) sparse grid and weights that come from nested quadrature rules (e.g. Clenshaw-Curtis).

In the UQ literature, these methods are discussed based on their error formulas, and the so-called *curse of dimensionality* is often mentioned. If $\boldsymbol{u}$ denotes the true expected value and $\boldsymbol{u}_N$ its approximation with N nodes/samples, then the methods behave in the following way:

- The MC error is determined by the root mean square error $E[(\boldsymbol{u} - \boldsymbol{u}_N)^2]^{1/2} = V_{MC}(\boldsymbol{u}) N^{-1/2}$.

- Multi-level Monte Carlo (MLMC) methods use control variates to make the constant $V(\boldsymbol{u})$ smaller [11].
- The QMC error typically behaves like $|\boldsymbol{u} - \boldsymbol{u}_N| \leq V_{QMC}(\boldsymbol{u})(\log N)^p N^{-1}$ [6].
- The tensorized grid error has the form $|\boldsymbol{u} - \boldsymbol{u}_N| \leq V_{\text{tens}}(\boldsymbol{u}) N^{-\alpha/p}$.
- The sparse grid quadrature error behaves like $|\boldsymbol{u} - \boldsymbol{u}_N| \leq V_{\text{sparse}}(\boldsymbol{u})(\log N)^p N^{-\beta}$.

In the latter two cases, α and β are related to the differentiability of $\boldsymbol{u}$ with respect to $\boldsymbol{\xi}$. All error constants V depend on $\boldsymbol{u}$ and certain of its derivatives. For instance, in the case of sparse grids, $\boldsymbol{u}$ has to be in a certain Sobolev space with mixed higher-order derivatives. For an overview and a critical discussion of the assumptions on the solution we refer the reader to the excellent paper [38]. In the mathematical UQ literature, the *curse of dimensionality* is usually defined as an effective decay of the convergence rate when it is measured in the total number of nodes N and when the dimension p is increased. This is clearly the case for tensorized grids. But one should also note that for both QMC and sparse grid quadrature the decaying term dominates the log term only if $N \gg 2^p$ ($N \gg 2^{p/\beta}$ respectively). These methods therefore only mitigate the curse. Finally, one often finds the statement that MC methods do not suffer from the curse, because the convergence rate is independent of the dimension p. However, MC methods might be impractical in high dimensions. This can be seen from the simple example (that every reader can easily try) of approximating the volume of the unit sphere in p dimensions: Draw a uniform sample in $[-1, 1]^p$ and determine if the sampled node is inside the unit sphere. Then the ratio of the points inside the sphere to the total number of samples converges to the volume divided by 2^p. For $p = 20$, MC with 100 million samples will not produce any significant digit. The reason is that the volume of the sphere becomes so small that it becomes almost impossible to draw a sample within the sphere. In other words, the error constant increases rapidly with dimension p. It should be noted, however, that all of these methods are embarrassingly parallel because uncoupled problems need to be solved.

Intrusive methods on the other hand do not rely on any form of quadrature, but rather derive a system of equations that describes the time evolution of the moments directly. The resulting system can then be solved with classical numerical methods for deterministic equations. In contrast to non-intrusive methods this does not decouple the problem. In the stochastic Galerkin (SG) method, the solution $\boldsymbol{u}$ is expanded in a series of polynomial basis functions $\varphi_i : \Theta \to \mathbb{R}$, such that for the multi-index $i = (i_1, \cdots, i_p)$ we have $|i| := \sum_{k=1}^{p} |i_k| \leq N$. The usual choice for these functions φ_i are orthonormal polynomials with respect to the probability distribution function, i.e. $\langle \varphi_i \varphi_j \rangle = \prod_{n=1}^{p} \delta_{i_n j_n}$. This yields the so called generalized polynomial chaos (gPC) expansion

$$\mathcal{U}(\hat{\boldsymbol{u}}; \boldsymbol{\xi}) := \sum_{|i| \leq N} \hat{\boldsymbol{u}}_i \varphi_i(\boldsymbol{\xi}) = \hat{\boldsymbol{u}}^T \boldsymbol{\varphi}(\boldsymbol{\xi}) \,. \tag{2}$$

The unknown, but deterministic expansion coefficients $\hat{\boldsymbol{u}}_i \in \mathbb{R}^m$ are called moments. For a more compact notation, we collect these moments in the moment matrix $\hat{\boldsymbol{u}}$. This matrix holds all moments for which $|i| \leq N$ holds. Therefore, $\hat{\boldsymbol{u}}$ is defined

as $\hat{\boldsymbol{u}} := (\hat{\boldsymbol{u}}_i)_{|i|\leq N} \in \mathbb{R}^{M\times m}$ with corresponding basis functions $\boldsymbol{\varphi} := (\varphi_i)_{|i|\leq N} \in \mathbb{R}^M$. The total number of basis functions for which $|i| \leq N$ holds is

$$M := \binom{N+p}{p} .$$

When the gPC approximation (2) is known, statistical quantities of interest can be computed as

$$\mathrm{E}[\mathcal{U}(\hat{\boldsymbol{u}})] = \hat{\boldsymbol{u}}_0 , \quad \mathrm{Var}[\mathcal{U}(\hat{\boldsymbol{u}})] = \mathrm{E}[\mathcal{U}(\hat{\boldsymbol{u}})^2] - \mathrm{E}[\mathcal{U}(\hat{\boldsymbol{u}})]^2 = \left(\sum_{i=1}^{N} \hat{u}_{\ell i}^2 \right)_{\ell=1,\cdots,m} .$$

The SG moment system is obtained by plugging the gPC ansatz (2) into the stochastic problem (1a) and projecting the resulting residual to zero (Galerkin projection), which yields the system

$$\partial_t \hat{\boldsymbol{u}}_i(t,\boldsymbol{x}) + \nabla \cdot \langle \boldsymbol{f}(\mathcal{U}(\hat{\boldsymbol{u}}))\varphi_i \rangle = \boldsymbol{0} , \tag{3a}$$

$$\hat{\boldsymbol{u}}_i(t=0,\boldsymbol{x}) = \langle \boldsymbol{u}_{\mathrm{IC}}(\boldsymbol{x})\varphi_i \rangle . \tag{3b}$$

As mentioned previously, the main caveat of the method is that the moment system is not necessarily hyperbolic and thus not applicable to every problem. We will investigate this thoroughly in the following sections, but at this point we want to discuss why one should be interested in an intrusive method like SG at all. Putting SG into practice requires working with the model and new code. Additionally, the trivial parallelism of non-intrusive methods is lost. A statement one often finds in the UQ literature is that one should use SG because it has spectral convergence. This means that the convergence rate of the method only depends on the smoothness of the function. Moreover, if this smoothness is large or even infinite (i.e. the solution possesses derivatives of orders up to infinity) then the curse of dimensionality can be overcome. However, both Gauss and Clenshaw-Curtis quadrature also show spectral convergence [39], so if a function is smooth enough those methods can be used as well.

On the plus side, because sparse grids rely on nested quadrature rules, and similar to modal versus nodal DG methods, SG reaches the same formal accuracy with fewer unknowns. Furthermore, in many cases the expected value of a solution of a hyperbolic system is more smooth and does not have shocks (examples can be found in Sect. 6). Although this is not true in general [37], one often does not have to use a high-resolution shock-capturing scheme. Two further advantages will be utilized in this paper: Whereas collocation methods use a global grid in the uncertain parameters, for intrusive methods one can enrich the discretization of the parameter space adaptively. Furthermore, especially for the IPM method one can iterate faster into steady state. Given these potential advantages we argue that although intrusive methods have shortcomings it is worthwhile to study them, especially in the context of hyperbolic conservation laws.

3 Hyperbolic Conservation Laws and the IPM Method

After introducing hyperbolic conservation laws and the concept of entropy, this section is focused on the derivation of the IPM method.

3.1 Hyperbolic Conservation Laws and Entropy Variables

Our numerical discretization of the random space should preserve certain properties of hyperbolic equations. Although they are well-known, we briefly summarize them in the following to fix notation. Ignoring uncertainties for the time being, to characterize hyperbolicity we put the conservative form (1a) into its *quasi-conservative form*. Defining the flux Jacobians $\boldsymbol{A}_j := \nabla_{\boldsymbol{u}} \boldsymbol{f}_j \in \mathbb{R}^{m\times m}$, the system (1a) can be rewritten as

$$\partial_t \boldsymbol{u} + \sum_{j=1}^{d} \boldsymbol{A}_j(\boldsymbol{u}) \partial_{x_j} \boldsymbol{u} = \boldsymbol{0} \,. \tag{4}$$

Denoting the flux Jacobian into direction $\boldsymbol{w} \in \mathbb{R}^d$ by

$$\boldsymbol{A}(\boldsymbol{u}, \boldsymbol{w}) := \sum_{j=1}^{d} \boldsymbol{A}_j w_j \,,$$

we call a system *hyperbolic*, if the flux Jacobian $\boldsymbol{A}(\boldsymbol{u}, \boldsymbol{w})$ has only real eigenvalues $\lambda_k(\boldsymbol{u}, \boldsymbol{w})$ for $k = 1, \cdots, m$ with a complete family of eigenvectors $\boldsymbol{r}_k(\boldsymbol{u}, \boldsymbol{w})$ for all states $\boldsymbol{u} \in \mathcal{D}$ and every direction $\boldsymbol{w} \in \mathbb{R}^d$ with $\|\boldsymbol{w}\| = 1$. Note that for one spatial dimension, i.e. $d = 1$, the direction is $w = 1$ and therefore hyperbolicity holds if the flux Jacobian $\nabla_{\boldsymbol{u}} \boldsymbol{f}$ is diagonalizable with real eigenvalues.

Hyperbolic problems tend to form shocks, in which case the original system can no longer be solved. Therefore, the concept of *weak solutions* has been introduced, which tests the original problem against smooth basis functions with compact support and then moves derivatives from the solution onto these basis functions [24, Chapter 3.4]. Unfortunately, weak solutions are not unique and can show non-physical behavior. Hence, one is left with having to pick physical meaningful solutions from possible weak solution candidates. This motivates a further concept, called the entropy solution [24, Chapter 3.8.1]. Note that in the case of scalar equations, the entropy solution is actually unique under certain smallness assumptions [16, Chapter 2.4]. Let us first introduce the entropy:

Definition 1 Let $\mathcal{D}$ be convex. Then a convex function $s : \mathcal{D} \to \mathbb{R}$ is called an *entropy* for the conservation Eqs. (1a) if there exist d functions $\tilde{F}_j : \mathcal{D} \to \mathbb{R}$, called *entropy fluxes*, which fulfill the *integrability condition*

$$\nabla_{\boldsymbol{u}} s(\boldsymbol{u}) \nabla_{\boldsymbol{u}} \boldsymbol{f}_j(\boldsymbol{u}) = \nabla_{\boldsymbol{u}} \tilde{F}_j(\boldsymbol{u}) \,, \quad j = 1, \cdots, d \,. \tag{5}$$

For classical solutions, the integrability condition ensures conservation of entropy: By multiplying $\nabla_{\boldsymbol{u}} s(\boldsymbol{u})$ from the left with the original Eq. (1a), we get with the integrability condition (5) as

$$\partial_t s(\boldsymbol{u}) + \sum_{j=0}^{d} \partial_{x_j} \tilde{F}_j(\boldsymbol{u}) = 0 \,. \tag{6}$$

Since (6) is again in conservation form, the entropy is conserved at smooth solutions and the functions $\tilde{F}_j$ are the flux functions of the entropy balance law (6).

If $\boldsymbol{u}$ is a weak solution, which fulfills

$$\partial_t s(\boldsymbol{u}) + \sum_{j=0}^{d} \partial_{x_j} \tilde{F}_j(\boldsymbol{u}) \leq 0 \tag{7}$$

in a weak sense for all admissible entropies s, then $\boldsymbol{u}$ is called an entropy solution. Note that opposed to the entropy used in thermodynamics, the mathematical entropy s is dissipated in time: By integrating (7) over the spatial domain, while assuming that the entropy fluxes are zero at the boundary, we obtain

$$\frac{d}{dt} \int_D s(\boldsymbol{u}) \, d\boldsymbol{x} \leq 0 \,. \tag{8}$$

The notion of entropy is closely related to hyperbolicity, which can be shown with the help of the *entropy variables*

$$\boldsymbol{v} = \nabla_{\boldsymbol{u}} s(\boldsymbol{u})^T \in \mathbb{R}^m \,. \tag{9}$$

If s is strictly convex, the mapping $\boldsymbol{v}(\boldsymbol{u})$ is one-to-one and the solution $\boldsymbol{u}$ can be represented in terms of entropy variables as $\boldsymbol{u} : \mathbb{R}^m \to \mathbb{R}^m$ with $\boldsymbol{u}(\boldsymbol{v}) = (\nabla_{\boldsymbol{u}} s)^{-1}(\boldsymbol{v})$.[1] A change from the conserved quantities $\boldsymbol{u}$ to their corresponding entropy variables can be performed to put (1a) in its *symmetric form*

$$\partial_t \boldsymbol{u}(\boldsymbol{v}) + \sum_{j=1}^{d} \partial_{x_j} \boldsymbol{g}_j(\boldsymbol{v}) = \boldsymbol{0} \,, \tag{10}$$

where the flux with respect to the entropy variables has been denoted by $\boldsymbol{g}_j$, i.e.

$$\boldsymbol{g}_j(\boldsymbol{v}) := \boldsymbol{f}_j(\boldsymbol{u}(\boldsymbol{v})) \qquad \text{with } j = 1, \cdots, d \,. \tag{11}$$

Our goal is to check hyperbolicity, i.e. (10) needs to be brought into its quasi-conservative form (4). Applying the chain rule results in

[1] Note that we have prescribed $\boldsymbol{u}$ to be in $\mathbb{R}^m$, i.e. strictly speaking we have $\boldsymbol{u}(\boldsymbol{v}) = (\nabla_{\boldsymbol{u}} s)^{-T}(\boldsymbol{v})$.

$$\boldsymbol{H}(\boldsymbol{v})\partial_t \boldsymbol{v} + \sum_{j=1}^{d} \boldsymbol{B}_j(\boldsymbol{v})\partial_{x_j} \boldsymbol{v} = \boldsymbol{0} \,, \tag{12}$$

with

$$\boldsymbol{H}(\boldsymbol{v}) = \nabla_{\boldsymbol{v}} \boldsymbol{u}(\boldsymbol{v}) \qquad \text{and} \qquad \boldsymbol{B}_j(\boldsymbol{v}) = \nabla_{\boldsymbol{v}} \boldsymbol{g}_j(\boldsymbol{v}) \,. \tag{13}$$

Note that $\boldsymbol{H}(\boldsymbol{v}) = \left(\nabla_{\boldsymbol{u}}^2 s(\boldsymbol{u})\right)^{-1}$, which can be checked by differentiating $\nabla_{\boldsymbol{u}} s(\boldsymbol{u}(\boldsymbol{v})) = \boldsymbol{v}$ with respect to $\boldsymbol{v}$. Therefore, $\boldsymbol{H}(\boldsymbol{v})$ is symmetric positive definite and can therefore be rewritten as $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$, where $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ is orthonormal and $\boldsymbol{\Lambda} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with positive entries. Consequently, the regular, symmetric matrix $\boldsymbol{H}^{1/2} := \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^T$ exists. Multiplying (12) with $\boldsymbol{H}^{-1} = \boldsymbol{H}^{-1/2}\boldsymbol{H}^{-1/2}$ from the left results in the system

$$\partial_t \boldsymbol{v} + \sum_{j=1}^{d} \boldsymbol{H}^{-1/2}\boldsymbol{H}^{-1/2}\nabla_{\boldsymbol{v}} \boldsymbol{g}_j(\boldsymbol{v})\boldsymbol{H}^{-1/2}\boldsymbol{H}^{1/2}\partial_{x_j} \boldsymbol{v} = \boldsymbol{0} \,. \tag{14}$$

It remains to check under which conditions the flux Jacobian of this system is diagonalizable with real eigenvalues. Since $\boldsymbol{H}^{-1/2}$ is symmetric, symmetry of $\boldsymbol{B}_j$ suffices to show symmetry of $\boldsymbol{H}^{-1/2}\nabla_{\boldsymbol{v}} \boldsymbol{g}_j(\boldsymbol{v})\boldsymbol{H}^{-1/2}$. Multiplying this matrix with $\boldsymbol{H}^{-1/2}$ from the left and $\boldsymbol{H}^{1/2}$ from the right is a similarity transformation and therefore does not change eigenvalues. Hence when $\boldsymbol{B}_j$ is symmetric, the system (14) is diagonalizable with real eigenvalues and therefore hyperbolic. This can be ensured via the concept of entropy.

Theorem 1 *The matrices $\boldsymbol{B}_j$ are symmetric iff the integrability condition (5) holds.*

Proof See e.g. [36]. □

In the case of scalar equations, all convex functions can be used as entropies. In particular, a family of entropies, which is also called the Kružkov entropy [18], given by

$$s(u) = |u - k| \qquad \text{for all } k \in \mathbb{R}$$

fulfills the integrability condition for the entropy flux

$$\tilde{F}_j(u) = \operatorname{sgn}(u - k)(f_j(u) - f_j(k)) \,.$$

This family of entropies can be employed to derive several solution properties for scalar equations. One of these properties is the *maximum–principle*

$$\|u(t,\cdot)\|_{L^\infty(D)} \leq \|u_{\mathrm{IC}}\|_{L^\infty(D)} \,, \tag{15}$$

which guarantees bounds on the solution imposed by its initial condition.

3.2 The Intrusive Polynomial Moment Method

Let us now consider a system of hyperbolic conservation laws of the form (1a) and move back to the discretization of the random domain. As discussed earlier, the polynomial ansatz of stochastic-Galerkin does not necessarily preserve hyperbolicity. A generalization of stochastic-Galerkin, which ensures hyperbolicity is the Intrusive Polynomial Moment (IPM) method [33], which in the field of kinetic theory is known as an entropy closure, see e.g. [25]. Instead of expanding the conserved variables $\boldsymbol{u}$ with polynomials as done by SG, the IPM method performs such an expansion on the entropy variables (9). Hence, substituting the entropy variables $\boldsymbol{v} = \nabla_{\boldsymbol{u}} s(\boldsymbol{u})^T$ into (1a) yields

$$\partial_t \boldsymbol{u}(\boldsymbol{v}(t,\boldsymbol{x},\boldsymbol{\xi})) + \nabla \cdot \boldsymbol{f}(\boldsymbol{u}(\boldsymbol{v}(t,\boldsymbol{x},\boldsymbol{\xi}))) = \boldsymbol{0}\,, \tag{16}$$

where again $\boldsymbol{u} : \mathbb{R}^m \to \mathbb{R}^m$ with $\boldsymbol{u}(\boldsymbol{v}) = \nabla_{\boldsymbol{u}} s(\boldsymbol{u})$. Now, a finite dimensional representation of the entropy variables is obtained by an expansion in terms of gPC polynomials, i.e.

$$\boldsymbol{v}(t,\boldsymbol{x},\boldsymbol{\xi}) \approx \boldsymbol{v}_N(t,\boldsymbol{x},\boldsymbol{\xi}) := \sum_{|i|\le N} \hat{\boldsymbol{v}}_i(t,\boldsymbol{x})\varphi_i(\boldsymbol{\xi}) = \hat{\boldsymbol{v}}(t,\boldsymbol{x})^T \boldsymbol{\varphi}(\boldsymbol{\xi})\,, \tag{17}$$

where the *entropic expansion coefficients* (also called *dual variables*) $\hat{\boldsymbol{v}}_i \in \mathbb{R}^m$ are collected in the matrix $\hat{\boldsymbol{v}} := (\hat{\boldsymbol{v}}_i)_{i\le|N|} \in \mathbb{R}^{M\times m}$. Replacing the exact entropy variables inside the original problem (16) by this expansion, we obtain

$$\partial_t \boldsymbol{u}\left(\hat{\boldsymbol{v}}(t,\boldsymbol{x})^T\boldsymbol{\varphi}(\boldsymbol{\xi})\right) + \nabla\cdot\boldsymbol{f}\left(\boldsymbol{u}\left(\hat{\boldsymbol{v}}(t,\boldsymbol{x})^T\boldsymbol{\varphi}(\boldsymbol{\xi})\right)\right) = \tilde{\boldsymbol{r}}(t,\boldsymbol{x},\boldsymbol{\xi})\,. \tag{18}$$

Similar to stochastic-Galerkin, the residual $\tilde{\boldsymbol{r}}$ is again projected to zero, yielding

$$\partial_t \left\langle \boldsymbol{u}\left(\hat{\boldsymbol{v}}(t,\boldsymbol{x})^T\boldsymbol{\varphi}\right)\varphi_i\right\rangle + \nabla\cdot\left\langle \boldsymbol{f}\left(\boldsymbol{u}\left(\hat{\boldsymbol{v}}(t,\boldsymbol{x})^T\boldsymbol{\varphi}\right)\right)\varphi_i\right\rangle = \boldsymbol{0} \tag{19}$$

for $|i| \le N$. The moments belonging to the dual variables $\hat{\boldsymbol{v}}$ are now given by

$$\hat{\boldsymbol{u}}_i(\hat{\boldsymbol{v}}) = \left\langle \boldsymbol{u}\left(\hat{\boldsymbol{v}}^T\boldsymbol{\varphi}\right)\varphi_i\right\rangle \qquad \text{for } |i| \le N\,. \tag{20}$$

This mapping, i.e. $\hat{\boldsymbol{u}} : \mathbb{R}^{M\times m} \to \mathcal{R} \subset \mathbb{R}^{M\times m}$ is one-to-one, meaning that similar to $\boldsymbol{v}(\boldsymbol{u})$, we can define a function $\hat{\boldsymbol{v}}(\hat{\boldsymbol{u}})$ with $\hat{\boldsymbol{v}} : \mathcal{R} \to \mathbb{R}^{M\times m}$. Making use of this mapping as well as the definition of the moments in (19) yields the IPM system

$$\partial_t \hat{\boldsymbol{u}}_i + \nabla\cdot\left\langle \boldsymbol{f}\left(\boldsymbol{u}\left(\hat{\boldsymbol{v}}(\hat{\boldsymbol{u}})^T\boldsymbol{\varphi}\right)\right)\varphi_i\right\rangle = \boldsymbol{0}\,, \quad \text{for } |i| \le N\,. \tag{21}$$

The IPM system posses several desirable properties. Especially, if the entropy $s(\boldsymbol{u})$ fulfills the integrability condition (5), the IPM system is hyperbolic:

Theorem 2 *The IPM system can be brought into its symmetric form with symmetric positive definite temporal Jacobian and symmetric spatial Jacobian, if the entropy $s(\boldsymbol{u})$ fulfills the integrability condition* (5).

Proof In its symmetric form, the IPM system (19) reads

$$\hat{\boldsymbol{H}}(\hat{\boldsymbol{v}})\partial_t \hat{\boldsymbol{v}} + \sum_{j=1}^{d} \hat{\boldsymbol{B}}_j(\hat{\boldsymbol{v}})\partial_{x_j}\hat{\boldsymbol{v}} = \boldsymbol{0}\,, \tag{22a}$$

with
$$\hat{\boldsymbol{H}}(\hat{\boldsymbol{v}}) := \left\langle \nabla_{\boldsymbol{v}}\boldsymbol{u}(\boldsymbol{v}_N) \otimes \boldsymbol{\varphi}\boldsymbol{\varphi}^T \right\rangle\,, \tag{22b}$$

$$\hat{\boldsymbol{B}}_j(\hat{\boldsymbol{v}}) := \left\langle \nabla_{\boldsymbol{u}}\boldsymbol{f}_j(\boldsymbol{u}(\boldsymbol{v}_N))\nabla_{\boldsymbol{v}}\boldsymbol{u}(\boldsymbol{v}_N) \otimes \boldsymbol{\varphi}\boldsymbol{\varphi}^T \right\rangle\,. \tag{22c}$$

Here, we abuse notation by defining the multiplication of $\hat{\boldsymbol{H}} \in \mathbb{R}^{m\cdot M\times m\cdot M}$ with $\boldsymbol{y} \in \mathbb{R}^{M\times m}$ by

$$\left(\hat{\boldsymbol{H}}\cdot\boldsymbol{y}\right)_{li} := \sum_{l'=1}^{m}\sum_{i'=1}^{M} \hat{H}_{(l-1)m+i,(l'-1)m+i'}\, y_{l'i'}\,.$$

The same holds for the multiplication with $\hat{\boldsymbol{B}}_j$. As done for (12), if we can ensure $\hat{\boldsymbol{H}}$ being symmetric positive definite and $\hat{\boldsymbol{B}}_j$ symmetric, we know that the IPM system is hyperbolic. Obviously, $\hat{\boldsymbol{H}}$ is symmetric. Multiplication with $\hat{\boldsymbol{v}} \in \mathbb{R}^{M\times m}$ from both sides gives

$$\hat{\boldsymbol{v}}^T\hat{\boldsymbol{H}}\hat{\boldsymbol{v}} = \left\langle \boldsymbol{v}_N^T\nabla_{\boldsymbol{v}}\boldsymbol{u}(\boldsymbol{v}_N)\boldsymbol{v}_N \right\rangle > 0\,,$$

where we use that $\nabla_{\boldsymbol{v}}\boldsymbol{u} = \boldsymbol{H}$ is symmetric positive definite as done in (12). It remains to show symmetry of $\hat{\boldsymbol{B}}_j$ for all $j = 1,\cdots,d$. Using the definition of $\boldsymbol{B}_j$ from (13), we can rewrite (22c) as

$$\hat{\boldsymbol{B}}_j(\hat{\boldsymbol{v}}) := \left\langle \boldsymbol{B}_j(\boldsymbol{v}_N) \otimes \boldsymbol{\varphi}\boldsymbol{\varphi}^T \right\rangle\,.$$

By Theorem 1, we know that $\boldsymbol{B}_j$ is symmetric, from which we can conclude symmetry of $\hat{\boldsymbol{B}}_j$. □

Recall that solving the IPM system requires the mapping $\hat{\boldsymbol{v}}(\hat{\boldsymbol{u}})$, i.e. a mapping from the moments to the dual variables. This mapping can be defined by inverting the dual variables to moments map (20). The inverse exists, since the Jacobian of $\hat{\boldsymbol{u}}(\hat{\boldsymbol{v}})$ is $\nabla_{\hat{\boldsymbol{v}}}\hat{\boldsymbol{u}}(\hat{\boldsymbol{v}}) = \hat{\boldsymbol{H}}(\hat{\boldsymbol{v}})$ which is positive definite, i.e. the dual variables to moments map is strictly monotonically increasing. Unfortunately, the inversion can generally not be performed analytically. In this case one needs to determine $\hat{\boldsymbol{v}}$ by solving the non-linear system of equations

$$\left\langle \boldsymbol{u}\left(\hat{\boldsymbol{v}}^T\boldsymbol{\varphi}\right)\boldsymbol{\varphi}^T \right\rangle^T = \hat{\boldsymbol{u}} \tag{23}$$

for a given moment vector $\hat{\boldsymbol{u}}$ numerically. This task is commonly performed by reformulating (23) as a root-finding problem

$$\mathcal{G}(\hat{\boldsymbol{v}};\hat{\boldsymbol{u}}) \stackrel{!}{=} \boldsymbol{0}$$

with

$$\mathcal{G}(\boldsymbol{w};\hat{\boldsymbol{u}}) := \left\langle \boldsymbol{u}\left(\boldsymbol{w}^T\boldsymbol{\varphi}\right)\boldsymbol{\varphi}^T\right\rangle^T - \hat{\boldsymbol{u}}\ . \tag{24}$$

Here, one often uses Newton's method to determine the root of $\mathcal{G}$. Then, with $\nabla_{\boldsymbol{w}}\mathcal{G}(\boldsymbol{w};\hat{\boldsymbol{u}}) = \hat{\boldsymbol{H}}(\boldsymbol{w})$ a Newton update takes the form $\boldsymbol{d} : \mathbb{R}^{M\times m}\times\mathbb{R}^{M\times m}\to\mathbb{R}^{M\times m}$ with

$$\boldsymbol{d}(\boldsymbol{w},\hat{\boldsymbol{u}}) := \boldsymbol{w} - \hat{\boldsymbol{H}}(\boldsymbol{w})^{-1}\cdot\mathcal{G}(\boldsymbol{w};\hat{\boldsymbol{u}})\ . \tag{25}$$

The function $\boldsymbol{d}$ will in the following be called dual iteration function. Now, the Newton iteration for an input moment vector $\hat{\boldsymbol{u}}$ is given by

$$\boldsymbol{w}^{(l+1)} = \boldsymbol{d}(\boldsymbol{w}^{(l)},\hat{\boldsymbol{u}})\ . \tag{26}$$

The exact dual state is then obtained by computing the fixed point of $\boldsymbol{d}$, meaning that one converges the iteration (26), i.e. $\hat{\boldsymbol{v}} := \hat{\boldsymbol{v}}(\hat{\boldsymbol{u}}) = \lim_{l\to\infty}\boldsymbol{d}(\boldsymbol{w}^{(l)},\hat{\boldsymbol{u}})$. To obtain a finite number of iterations, a stopping criterion

$$\sum_{i=0}^{m}\left\|\mathcal{G}(\boldsymbol{w}^{(l)};\hat{\boldsymbol{u}})\right\| < \tau \tag{27}$$

is used, where $\tau > 0$ is a user determined parameter.

It remains to discuss the discretization of the spatial and time domain, which for ease of presentation, we perform for a scalar problem as well as a one dimensional spatial domain. When dividing the spatial domain into cells $[x_{j-1/2}, x_{j+1/2}]$ with $j = 1,\cdots,N_x$ and using discrete times t_n with $n = 1,\cdots,N_t$, we can approximate the i-th order moment by

$$\hat{\boldsymbol{u}}_{ij}^n \simeq \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}}\hat{\boldsymbol{u}}_i(t_n,x)dx\ .$$

The full moment vector in cell j at time t_n is denoted by $\hat{\boldsymbol{u}}_j^n = (\hat{\boldsymbol{u}}_{0j}^n,\cdots,\hat{\boldsymbol{u}}_{Nj}^n)^T \in \mathbb{R}^{M\times m}$. Furthermore, the corresponding dual variables are denoted by

$$\hat{\boldsymbol{v}}_j^n := \hat{\boldsymbol{v}}\left(\hat{\boldsymbol{u}}_j^n\right)\ . \tag{28}$$

Then, a finite-volume scheme for the IPM system (21) can be written as

$$\hat{\boldsymbol{u}}_j^{n+1} = \hat{\boldsymbol{u}}_j^n - \frac{\Delta t}{\Delta x}\left(\boldsymbol{G}^*(\hat{\boldsymbol{v}}_j,\hat{\boldsymbol{v}}_{j+1}) - \boldsymbol{G}^*(\hat{\boldsymbol{v}}_{j-1},\hat{\boldsymbol{v}}_j)\right)\ , \tag{29}$$

where $\boldsymbol{G}^* : \mathbb{R}^{M\times m}\times\mathbb{R}^{M\times m}\to\mathbb{R}^{M\times m}$ is the numerical flux which needs to be consistent with the physical flux of the IPM system. i.e. we must have $\boldsymbol{G}^*(\hat{\boldsymbol{v}},\hat{\boldsymbol{v}}) = \left\langle \boldsymbol{f}\left(\boldsymbol{u}\left(\hat{\boldsymbol{v}}^T\boldsymbol{\varphi}\right)\right)\boldsymbol{\varphi}^T\right\rangle$. In order to evaluate the moments to dual variables map (28), we

can use the defined Newton iteration for the input moments $\hat{\boldsymbol{u}}_j^n$. Altogether, this yields the scheme from Algorithm 1.

Algorithm 1 IPM algorithm

for $j = 0$ to $N_x + 1$ **do**
  $\hat{\boldsymbol{u}}_j^0 = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \langle u_{\mathrm{IC}}(x, \cdot)\boldsymbol{\varphi} \rangle dx$
for $n = 0$ to N_t **do**
  for $j = 0$ to $N_x + 1$ **do**
    while $\|\mathcal{G}(\boldsymbol{v}_j^{(m)}; \hat{\boldsymbol{u}}_j^n)\| > \tau$ **do**
      $\boldsymbol{v}_j^{(m+1)} \leftarrow \boldsymbol{d}(\boldsymbol{v}_j^{(m)}; \hat{\boldsymbol{u}}_j^n)$
      $m \leftarrow m + 1$
    $\hat{\boldsymbol{v}}_j^n \leftarrow \boldsymbol{v}_j^{(m+1)}$
  for $j = 1$ to N_x **do**
    $\hat{\boldsymbol{u}}_j^{n+1} \leftarrow \hat{\boldsymbol{u}}_j^n - \frac{\Delta t}{\Delta x} \left[\boldsymbol{G}^*(\hat{\boldsymbol{v}}_j, \hat{\boldsymbol{v}}_{j+1}) - \boldsymbol{G}^*(\hat{\boldsymbol{v}}_{j-1}, \hat{\boldsymbol{v}}_j) \right]$

4 Realizability-Preserving Spatial Discretization

In this section we further describe the concept of realizability and present a realizability-preserving discretization and improved version of Algorithm 1.

4.1 Realizability

As previously discussed, the IPM method and minimal entropy closures in general face several challenges. Besides increased computational costs, the IPM method cannot invert the mapping $\hat{\boldsymbol{u}} : \mathbb{R}^{M \times m} \to \mathcal{R} \subset \mathbb{R}^{M \times m}$ when the moment vector $\hat{\boldsymbol{u}}$ leaves the so-called realizable set $\mathcal{R}$, which results in a failure of the method [19]. To discuss this issue, we consider a scalar, one-dimensional conservation law of the form

$$\partial_t u(t, x, \xi) + \partial_x f(u(t, x, \xi)) = 0 \,, \tag{30a}$$

$$u(0, x, \xi) = u_{\mathrm{IC}}(x, \xi) \,, \tag{30b}$$

i.e. m, p and d are equal to one. The following discussion is however valid for arbitrary dimensions and we make this simplification for ease of exposition. For scalar problems of the form (30), the solution fulfills the maximum–principle [16, Chapter 2.4]

$$\min_{x \in D, \xi \in \Theta} u_{\mathrm{IC}}(x, \xi) \leq u(t, x, \xi) \leq \max_{x \in D, \xi \in \Theta} u_{\mathrm{IC}}(x, \xi) \,,$$

which ideally should be preserved by the discretization of the random domain. The IPM method at least enables one to impose a user-defined lower bound u_- and upper bound u_+ on the solution by choosing an entropy $s(u)$, which takes infinite values when $u \notin (u_-, u_+)$. One such entropy is the log-barrier entropy [33]

$$s(u) = -\ln(u - u_-) - \ln(u_+ - u) \ . \tag{31}$$

Then, by the entropy dissipation property (8), the IPM solution $u(v_N) \equiv (s')^{-1}(v_N)$ will remain inside the interval (u_-, u_+). Similarly, for systems such as the Euler equations, certain solution quantities such as positivity of density, energy and pressure can again be achieved by the choice of a suitable entropy. Recall, that the image of the dual variables to moments map (20) has been denoted by $\mathcal{R}$. This set is called *realizable set*. For entropies imposing solution bounds u_- and u_+ it is given by

$$\mathcal{R} := \left\{ \hat{\boldsymbol{u}} \in \mathbb{R}^{N+1} \middle| \, \exists u : \Theta \to (u_-, u_+) \text{ such that } \hat{\boldsymbol{u}} = \langle u \boldsymbol{\varphi} \rangle \right\} \ . \tag{32}$$

When proposing numerical methods to solve the IPM system (21), it is crucial to prevent the moments generated by this method from leaving this set, since then, the dual variables to moments map cannot be inverted, i.e. the system (23) has no solution. In the following, we propose an algorithm which keeps the moments inside $\mathcal{R}$ and we will refer to this property as preserving realizability.

4.2 Realizability-Preserving Discretization

The presented general Algorithm 1 will not necessarily preserve realizability, i.e. it generates moments $\hat{\boldsymbol{u}} \notin \mathcal{R}$. The two sources for this are the choice of the numerical flux as well as the fact that the system (23) cannot be solved exactly, i.e. the moments to dual variables map has errors. Let us first write down the realizability preserving algorithm presented in [19] and then discuss why it maintains $\hat{\boldsymbol{u}} \notin \mathcal{R}$. When again using $u : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ with $u(v) = (s')^{-1}(v)$, the chosen numerical flux is the *kinetic flux*

$$\boldsymbol{G}^*(\hat{\boldsymbol{v}}_\ell, \hat{\boldsymbol{v}}_r) = \left\langle f^*(u(\hat{\boldsymbol{v}}_\ell^T \boldsymbol{\varphi}), u(\hat{\boldsymbol{v}}_r^T \boldsymbol{\varphi})) \boldsymbol{\varphi} \right\rangle \ , \tag{33}$$

where $f^*(u_\ell, u_r)$ is a monotone flux for the underlying deterministic problem. Note that this choice of $\boldsymbol{G}^*$ is common in the field of kinetic theory, see e.g. [8, 13, 31, 32]. We assume that the original, deterministic scheme

$$H(u, v, w) = v - \frac{\Delta t}{\Delta x} \left(f^*(v, w) - f^*(u, v) \right) \tag{34}$$

keeps the solution inside the bounds u_- and u_+, i.e. $H(u_{j-1}^n, u_j^n, u_{j+1}^n) \in [u_-, u_+]$ if all inputs are bounded by u_-, u_+. This can for example be achieved with monotone schemes or, for high order methods, with bound preserving limiters [4, 7, 12, 26,

41]. Note that since the integral in (33) can generally not be solved analytically, a quadrature rule must be employed to approximate the kinetic flux. It can however be shown that the resulting quadrature error will not influence realizability of the numerical scheme. Then, a realizability preserving implementation is given by

Algorithm 2 Modified IPM algorithm

1: **for** $j = 0$ to $N_x + 1$ **do**
2: $\quad \hat{\boldsymbol{u}}_j^0 = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \langle u_{\mathrm{IC}}(x, \cdot)\boldsymbol{\varphi} \rangle dx$
3: **for** $n = 0$ to N_t **do**
4: $\quad$ **for** $j = 0$ to $N_x + 1$ **do**
5: $\quad\quad$ **while** $\|\mathcal{G}(\boldsymbol{v}_j^{(m)}; \hat{\boldsymbol{u}}_j^n)\| > \tau$ **do**
6: $\quad\quad\quad \boldsymbol{v}_j^{(m+1)} \leftarrow \boldsymbol{d}(\boldsymbol{v}_j^{(m)}; \hat{\boldsymbol{u}}_j^n)$
7: $\quad\quad\quad m \leftarrow m + 1$
8: $\quad\quad \bar{\boldsymbol{v}}_j^n \leftarrow \boldsymbol{v}_j^{(m+1)}$
9: $\quad\quad \bar{\boldsymbol{u}}_j^n \leftarrow \left\langle u\left(\left(\bar{\boldsymbol{v}}_j^n\right)^T \boldsymbol{\varphi}\right)\boldsymbol{\varphi} \right\rangle$
10: $\quad$ **for** $j = 1$ to N_x **do**
11: $\quad\quad \hat{\boldsymbol{u}}_j^{n+1} \leftarrow \bar{\boldsymbol{u}}_j^n - \frac{\Delta t}{\Delta x}\left[\boldsymbol{G}^*(\bar{\boldsymbol{v}}_j, \bar{\boldsymbol{v}}_{j+1}) - \boldsymbol{G}^*(\bar{\boldsymbol{v}}_{j-1}, \bar{\boldsymbol{v}}_j)\right]$

Here, we use $\bar{\boldsymbol{v}}$ and $\bar{\boldsymbol{u}}$ instead of $\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{u}}$ to stress that these quantities are affected by the inexact Newton iteration. The main difference to Algorithm 1, besides the choice of the numerical flux, is the recalculation of the moment vector from the inexact dual variables $\bar{\boldsymbol{v}}$ in line 9. It can be shown that Algorithm 2 preserves realizability: To simplify notation, let us define the exact and inexact dual states $\Lambda := v_N = \hat{\boldsymbol{v}}^T\boldsymbol{\varphi}$ and $\overline{\Lambda} := \bar{\boldsymbol{v}}^T\boldsymbol{\varphi}$. Then, the inexact moments are given by

$$\bar{\boldsymbol{u}}_j^n = \left\langle u\left(\overline{\Lambda}_j^n\right)\boldsymbol{\varphi} \right\rangle \tag{35}$$

and the moment update becomes

$$\hat{\boldsymbol{u}}_j^{n+1} = \bar{\boldsymbol{u}}_j^n - \frac{\Delta t}{\Delta x}\left[\left\langle f^*\left(u\left(\overline{\Lambda}_{j-1}^n\right), u\left(\overline{\Lambda}_j^n\right)\right)\boldsymbol{\varphi}\right\rangle - \left\langle f^*\left(u\left(\overline{\Lambda}_j^n\right), u\left(\overline{\Lambda}_{j+1}^n\right)\right)\boldsymbol{\varphi}\right\rangle\right] . \tag{36}$$

Plugging the definition of the inexact moments (35) into the moment update (36) yields

$$\begin{aligned}\hat{\boldsymbol{u}}_j^{n+1} &= \left\langle \left(u\left(\overline{\Lambda}_j^n\right) - \frac{\Delta t}{\Delta x}\left[f^*\left(u\left(\overline{\Lambda}_j^n\right), u\left(\overline{\Lambda}_{j+1}^n\right)\right) - f^*\left(u\left(\overline{\Lambda}_{j-1}^n\right), u\left(\overline{\Lambda}_j^n\right)\right)\right]\right)\boldsymbol{\varphi} \right\rangle \\ &= \left\langle H(u(\overline{\Lambda}_{j-1}^n), u(\overline{\Lambda}_j^n), u(\overline{\Lambda}_{j+1}^n))\boldsymbol{\varphi} \right\rangle .\end{aligned}$$

Now, since the ansatz $u(\Lambda) = (s')^{-1}(\Lambda)$ only takes values in $[u_-, u_+]$, we have

$$H(u(\overline{\Lambda}_{j-1}^n), u(\overline{\Lambda}_j^n), u(\overline{\Lambda}_{j+1}^n)) \in [u_-, u_+]$$

for all $\xi \in \Theta$. Therefore, the time updated moments belong to an underlying function which is bounded by u_- and u_+, i.e. we have $\hat{\boldsymbol{u}}_j^{n+1} \in \mathcal{R}$. The construction of the presented algorithm also guarantees that the CFL condition of the original scheme ensures stability. Note that the presented IPM scheme can be extended to systems, multi-dimensional problems and higher order methods, whenever a bound-preserving scheme (34) exists. Furthermore, when replacing all integrals by quadrature rules, the time updated moments will remain realizable, since they belong to a function which fulfills the prescribed bounds on the quadrature points. The main drawback of this strategy to preserve realizability is that it introduces a non-conservative error by recalculating moments. However, when using a Lipschitz continuous numerical flux $\boldsymbol{G}^*$, the error by recalculating moments is of order $\mathcal{O}(\tau)$ [19]. I.e. by choosing a sufficiently small stopping criterion for Newton's method the error from the recalculation step becomes negligibly small.

A second strategy, which does not require recomputing moments and therefore does not add such an error is choosing a modified CFL condition. Here, the main idea is to account for effects the error in Λ has on the scheme by choosing a smaller time step size. Denoting this error by

$$\Delta\Lambda_j^n = \Delta\Lambda_j^n(\xi) := \Lambda_j^n(\xi) - \overline{\Lambda}_j^n(\xi) \, ,$$

a more restrictive CFL condition, which ensures realizability is given by

Theorem 3 *Let us assume that the entropy ansatz only takes values in (u_-, u_+) and the underlying numerical flux f^* is monotone. If furthermore, the numerical optimizer enforces the stopping criterion*

$$\max_{\xi\in\Theta} \left\{ \max_{\Lambda\in\left[\bar{\Lambda}_{j,min}^n, \bar{\Lambda}_{j,max}^n\right]} \frac{u'(\Lambda(\xi))}{u'(\Lambda(\xi) + \Delta\Lambda_j^n(\xi))} \right\} \leq \gamma \, , \tag{37}$$

with

$$\begin{aligned} \overline{\Lambda}_{j,min}^n(\xi) &:= \min\left\{\overline{\Lambda}_{j-1}^n(\xi), \overline{\Lambda}_j^n(\xi), \overline{\Lambda}_{j+1}^n(\xi)\right\} \\ \text{and} \quad \overline{\Lambda}_{j,max}^n(\xi) &:= \max\left\{\overline{\Lambda}_{j-1}^n(\xi), \overline{\Lambda}_j^n(\xi), \overline{\Lambda}_{j+1}^n(\xi)\right\} \, , \end{aligned}$$

the time updated moment vector $\hat{\boldsymbol{u}}_j^{n+1}$ is realizable under the modified CFL condition

$$\gamma \frac{\Delta t}{\Delta x} \max_{u\in[u_-,u_+]} |f'(u)| \leq 1 \, . \tag{38}$$

The proof of this theorem as well as an implementation strategy can be found in [19]. Due to its simplicity and efficient implementation of different discretization schemes, we will use the strategy of recalculating moments, i.e. Algorithm 2 in the following.

5 Accelerating the IPM Solution

In this section we describe two acceleration techniques, as presented in [23], which reduce the numerical effort of the IPM method.

5.1 Adaptivity

As previously mentioned, in contrast to non-intrusive methods, intrusive methods allow a more fine-grained control over the solution, as the uncertainties or more precisely their respective moments, are directly propagated through time and the corresponding quantities of interest (e.g. mean or variance) are not collected in a secondary step as in e.g. SC or MC methods. Using adaptivity, we try to avoid using high-order moment representations and corresponding high order quadrature rules in portions of the domain, where the quantities of interest are well-represented with low-order moments. As these lower-order moment bases result in non-linear system (23) that are easier to solve, this approach can significantly reduce overall runtimes. We use the discontinuity sensor described in [30] in the UQ context. To do this, the polynomial approximation at refinement level ℓ is defined as

$$\tilde{\boldsymbol{u}}_\ell := \sum_{|i| \leq M_\ell} \boldsymbol{u}_i \varphi_i \ .$$

We further define an indicator for a moment vector at level ℓ as

$$\boldsymbol{S}_\ell := \frac{\langle \left(\tilde{\boldsymbol{u}}_\ell - \tilde{\boldsymbol{u}}_{\ell-1} \right)^2 \rangle}{\langle \tilde{\boldsymbol{u}}_\ell^2 \rangle} \ . \tag{39}$$

Note, that a similar indicator has been used in [17] for intrusive methods in UQ. We use the first element in $\boldsymbol{S}_\ell$, i.e. the density ρ, to determine the refinement level. This regularity indicator is therefore computed for every cell at every timestep and the current refinement level is kept if the indicator lies in the interval $I_\delta := [\delta_-, \delta_+]$. If its value falls below δ_-, the refinement level is decreased to the next lower refinement level and vice versa if the value exceeds δ_+. See [23] for more details on the method.

5.2 One-Shot IPM

The second method is limited to steady state problems. In this case, we are interested in solving

$$\nabla \cdot \boldsymbol{f}(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\xi})) = \boldsymbol{0} \quad \text{in } D \tag{40}$$

with adequate boundary conditions. Then, the IPM moment system reads

$$\nabla \cdot \langle \boldsymbol{f}(\boldsymbol{u}(\boldsymbol{v}_N(\boldsymbol{x}, \boldsymbol{\xi})))\boldsymbol{\varphi}^T \rangle^T = \boldsymbol{0} \quad \text{in } D \ . \tag{41}$$

Steady state problems are usually solved by introducing a pseudo-time and iterating the solution until the condition

$$\sum_{j=1}^{N_x} \Delta x_j \| \hat{\boldsymbol{u}}_j^n - \hat{\boldsymbol{u}}_j^{n-1} \| \leq \varepsilon \ , \tag{42}$$

with convergence tolerance ε, is fulfilled. To obtain a more compact notation, let us define the pseudo-time update of the moments by

$$\begin{aligned} \boldsymbol{c}(\boldsymbol{w}_\ell, \boldsymbol{w}_c, \boldsymbol{w}_r) &:= \langle \boldsymbol{u}(\boldsymbol{w}_c^T \boldsymbol{\varphi})\boldsymbol{\varphi}^T \rangle^T \\ &\quad - \frac{\Delta t}{\Delta x} \left(\langle \boldsymbol{g}(\boldsymbol{u}(\boldsymbol{w}_c^T \boldsymbol{\varphi}), \boldsymbol{u}(\boldsymbol{w}_r^T \boldsymbol{\varphi}))\boldsymbol{\varphi}^T \rangle^T - \langle \boldsymbol{g}(\boldsymbol{u}(\boldsymbol{w}_\ell^T \boldsymbol{\varphi}), \boldsymbol{u}(\boldsymbol{w}_c^T \boldsymbol{\varphi}))\boldsymbol{\varphi}^T \rangle^T \right) \ . \end{aligned} \tag{43}$$

Note that the first term of this update recalculates moments from inexact dual variables, i.e. we perform a recalculation step according to Algorithm 2. To indicate the usage of inexact dual states, we again use the notation $\bar{\boldsymbol{v}}_j^n$. Then, the moment iteration of cell j, which is performed until (42) is fulfilled reads

$$\hat{\boldsymbol{u}}_j^{n+1} = \boldsymbol{c}\left(\bar{\boldsymbol{v}}_{j-1}^n, \bar{\boldsymbol{v}}_j^n, \bar{\boldsymbol{v}}_{j+1}^n\right) \ . \tag{44}$$

During each iteration, the dual variables $\bar{\boldsymbol{v}}_j^n$ are again obtained by iterating

$$\boldsymbol{v}_j^{(l+1)} = \boldsymbol{d}(\boldsymbol{v}_j^{(l)}; \hat{\boldsymbol{u}}_j^n)$$

until the stopping criterion (27) is fulfilled. A schematic of this method is given in Fig. 1. In the following, we refer to updating the dual variables as the inner loop and the iteration of the moments as the outer loop. The key idea of the *One-Shot IPM* (osIPM) method is to break up the inner loop and iterate moments and dual variables to their steady state simultaneously. This method is motivated by the One-Shot method in shape optimization [15], which proposes to perform only a single iteration for the primal, dual and design updates. The osIPM method now reads

$$\boldsymbol{v}_j^{n+1} = \boldsymbol{d}(\boldsymbol{v}_j^n, \hat{\boldsymbol{u}}_j^n) \quad \text{for all j} \ , \tag{45a}$$

$$\hat{\boldsymbol{u}}_j^{n+1} = \bar{\boldsymbol{u}}_j^n - \frac{\Delta t}{\Delta x} \left[\boldsymbol{G}^*(\boldsymbol{v}_{j-1}, \boldsymbol{v}_j) - \boldsymbol{G}^*(\boldsymbol{v}_j, \boldsymbol{v}_{j+1}) \right] \quad \text{for all j} \ . \tag{45b}$$

Note that the dual variables from the One-Shot iteration are written without a bar to indicate that they are not intended to be a solution of the dual problem. It can be

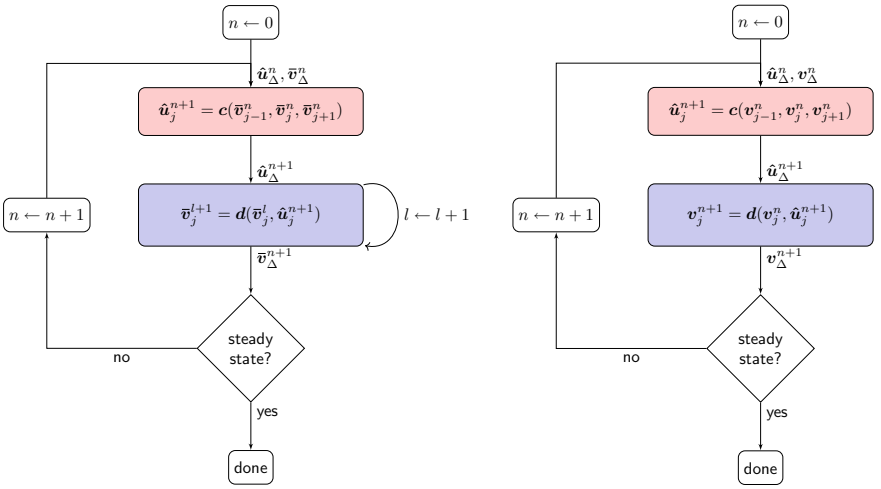


Fig. 1 *Left:* IPM method for steady state problems. *Right:* osIPM method. The use of Δ indicates that all spatial cells of the corresponding quantity are collected in a vector.

shown that the osIPM method converges locally [23]. Numerical studies show that the One-Shot IPM method requires more iterations of the outer loop compared to the general IPM method, but as these iterations are significantly cheaper in terms of computational effort, the method yields a significant boost in the performance [23].

6 Results

In order to demonstrate the properties of the presented IPM method and the related acceleration techniques, in this section we show four different test cases for the Burgers and the Euler equations, each highlighting different aspects.

6.1 Burgers' Equation

In the following, we investigate Burgers' forming shock testcase from [33], which has also been investigated in [19–22]. The stochastic Burgers equation for a one-dimensional spatial domain is given by

$$\partial_t u(t,x,\xi) + \partial_x \frac{u(t,x,\xi)^2}{2} = 0\,,$$
$$u(t=0,x,\xi) = u_{\mathrm{IC}}(x,\xi)\,.$$

The initial condition is now randomly distributed, i.e. we have

$$u_{\mathrm{IC}}(x,\xi) := \begin{cases} u_L, & \text{if } x < x_0 + \sigma\xi \\ u_L + \frac{u_R - u_L}{x_0 - x_1}(x_0 + \sigma\xi - x), & \text{if } x \in [x_0 + \sigma\xi, x_1 + \sigma\xi] \\ u_R, & \text{else} \end{cases} . \quad (46)$$

The initial condition describes a forming shock with a linear connection from $x_0 + \sigma\xi$ to $x_1 + \sigma\xi$, i.e. the random variable ξ shifts the forming shock structure to the left and right. We assume a uniformly distributed random variable in the interval $[-1, 1]$, i.e. $\xi \sim U([-1, 1])$. Furthermore, we use the following parameter values:

$D = [a, b] = [0, 3]$	Range of spatial domain
$N_x = 2000$	Number of spatial cells
$t_{\mathrm{end}} = 0.0909$	End time
$x_0 = 0.5, x_1 = 1.5, u_L = 12, u_R = 1, \sigma = 0.2$	Parameters of initial condition (46)
$N + 1 = 6$	Number of moments
$\varepsilon = 10^{-7}$	Accuracy of Newton's method

We compare the solution in ξ at a fixed spatial position x^* for time t_{end} for stochastic-Galerkin and IPM in Fig. 2. The IPM method uses the *bounded–barrier* entropy

$$s(u) = (u - u_-)\ln(u - u_-) + (u_+ - u)\ln(u_+ - u) ,$$

which, in contrast to the log–barrier entropy (31) takes finite values at u_- and u_+. Indeed, as seen in Sect. 4.2, it suffices that the ansatz $\left(s'\right)^{-1}$ only takes values in $[u_-, u_+]$ to enforce such bounds. For the bounded–barrier entropy, we can choose the distance to the exact solution to be zero, i.e. we have $\Delta u := u_R - u_- = u_+ - u_L = 0$. We also show results for the log–barrier entropy with $\Delta u = 0.5$. It can be seen that stochastic-Galerkin oscillates heavily while both IPM solutions maintain the overall shock characteristics. However, the bounded–barrier entropy is able to capture the shock more adequately while maintaining the maximum–principle (15). In the following, we focus on the bounded–barrier entropy and investigate its behavior when approximating expectation value and variance. For this, we let the simulation run until an increased end time $t_{\mathrm{end}} = 0.14$ is reached. Expectation value and variance are shown in Fig. 3. While stochastic-Galerkin yields a step-like profile, the IPM method when using the bounded–barrier entropy shows a significantly improved solution. Note, that the log–barrier entropy yields a similar result.

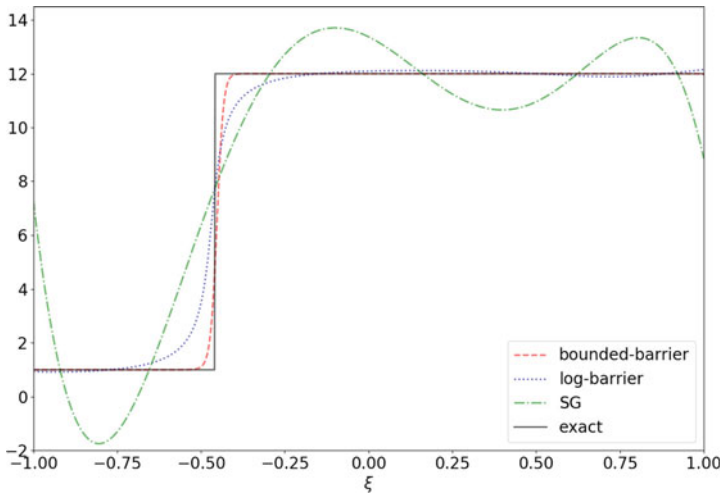


Fig. 2 Solution SG and IPM with bounded–barrier ($\Delta u = 0$) and log–barrier ($\Delta u = 0.5$) entropies at fixed spatial position $x^* = 1.5$ for time $t_{\text{end}} = 0.0909$.

6.2 *Euler Equations*

A further commonly used test case for intrusive methods is Sod's shock tube with uncertain shock position, see for example [22, 33, 34]. The stochastic Euler equations in one spatial dimension read

$$\partial_t \begin{pmatrix} \rho \\ \rho u \\ \rho e \end{pmatrix} + \partial_x \begin{pmatrix} \rho u \\ \rho u^2 + p \\ u(\rho e + p) \end{pmatrix} = \mathbf{0} \,,$$

where, in our test case, we use the initial condition

$$\begin{aligned} \rho_{\text{IC}} &= \begin{cases} \rho_L & \text{if } x < x_{\text{interface}}(\xi) \\ \rho_R & \text{else} \end{cases} \,, \\ (\rho u)_{\text{IC}} &= 0 \,, \\ (\rho e)_{\text{IC}} &= \begin{cases} \rho_L e_L & \text{if } x < x_{\text{interface}}(\xi) \\ \rho_R e_R & \text{else} \end{cases} \,. \end{aligned}$$

Here, ρ denotes the density, u is the velocity and e is the specific total energy. One can determine the pressure p from

$$p = (\gamma - 1)\rho \left(e - \frac{1}{2} u^2 \right) \,.$$

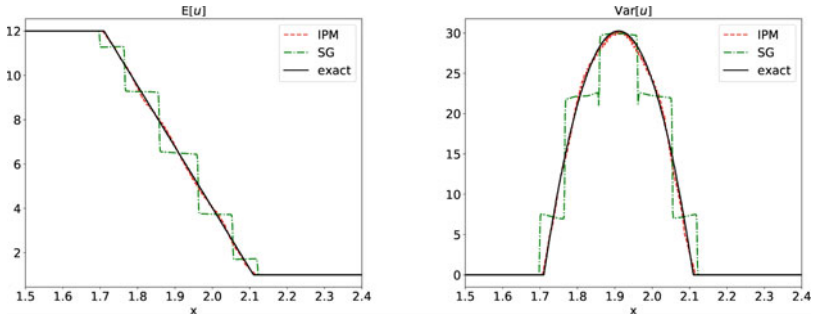


Fig. 3 Expectation value and variance for SG and IPM at time $t_{\text{end}} = 0.14$.

The heat capacity ratio is γ and has a value of 1.4 for air. We use the random interface position $x_{\text{interface}}(\xi) = x_0 + \sigma\xi$, where ξ is again uniformly distributed in the interval $[-1, 1]$. The IPM method again needs to pick a suitable entropy. In this work, we choose the entropy

$$s(\rho, \rho u, \rho e) = -\rho \ln\left(\rho^{-\gamma}\left(\rho e - \frac{(\rho u)^2}{2\rho}\right)\right),$$

though more choices are possible. Parameter values which differ from Burgers' test case are

$D = [a, b] = [0, 1]$	Range of spatial domain
$N_x = 5000$	Number of spatial cells
$t_{\text{end}} = 0.14$	End time
$x_0 = 0.5, \sigma = 0.05$	Interface position parameters
$\rho_L = 1.0, e_L = 2.5, \rho_R = 0.125, e_R = 0.25$	Initial states

When running the simulation, the SG method fails already during the first time update. The reason for this can be seen in Fig. 4. Here, the SG and IPM reconstructions of the gas density ρ are depicted at $t = 0$ s at a fixed spatial cell. While the IPM reconstruction maintains positivity, the Gibbs phenomena that result from the polynomial representation of SG lead to negative density values. A similar behavior can be seen for the energy e. Then, the eigenvalues of the Euler equations, which include $v \pm \sqrt{\gamma p/\rho}$ become complex, i.e. the system is no longer hyperbolic.

As discussed, the IPM method maintains hyperbolicity, meaning that one can run the simulation until the desired end time $t_{\text{end}} = 0.14$ s is reached. The resulting expectation values and variances are depicted in Fig. 5. It can be seen that the IPM method yields a satisfactory approximation of the expectation value and variance at the rarefaction wave as well as the contact discontinuity. However, the shock

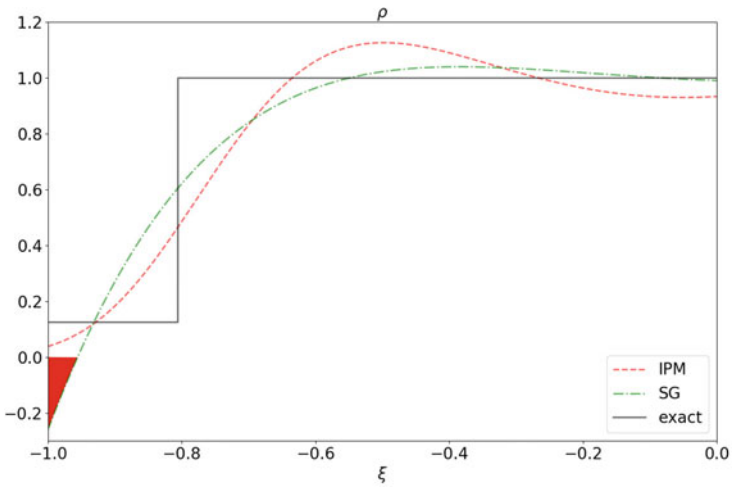


Fig. 4 Initial density for SG and IPM at fixed spatial position $x^* = 0.46$ when using 6 moments. The view is zoomed to $\xi \in [-1, 0]$ and negative regions are marked in red.

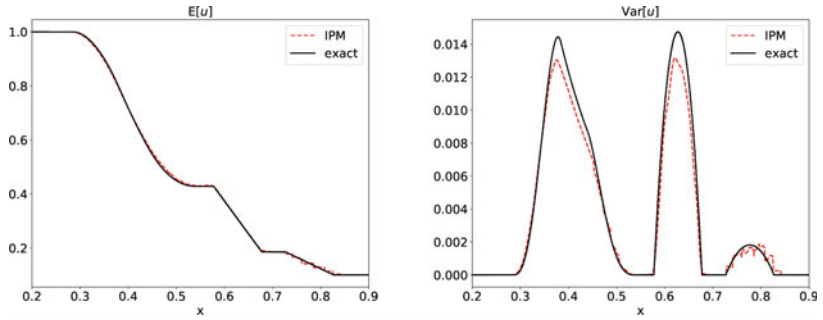


Fig. 5 Expectation value and variance for SG and IPM at time $t_{\text{end}} = 0.14$ s.

yields discontinuous step-like profiles, similar to the stochastic-Galerkin results for Burgers' equation.

6.3 2-D Euler Equations with One-Shot

In order to demonstrate the acceleration impact of the aforementioned One-Shot strategy for the steady-state case, we will quantify the effects of an uncertain angle of attack $\phi \sim U(0.75, 1.75)$ for a NACA0012 airfoil using the Euler equations in two spatial dimensions. This test-case is taken from [23]. Similar to the 1-D case, the stochastic Euler equations in two dimensions are given by

$$\partial_t \begin{pmatrix} \rho \\ \rho u_1 \\ \rho u_2 \\ \rho e \end{pmatrix} + \partial_{x_1} \begin{pmatrix} \rho u_1 \\ \rho u_1^2 + p \\ \rho u_1 u_2 \\ u_1(\rho e + p) \end{pmatrix} + \partial_{x_2} \begin{pmatrix} \rho u_2 \\ \rho u_1 u_2 \\ \rho u_2^2 + p \\ u_2(\rho e + p) \end{pmatrix} = \mathbf{0} \,,$$

with the closure term for the pressure as

$$p = (\gamma - 1)\rho \left(e - \frac{1}{2}(u_1^2 + u_2^2) \right) \,.$$

As in the previous test case, the heat capacity ratio is set to $\gamma = 1.4$. For this test case, we apply the Euler slip boundary condition to the airfoil's boundary as $\boldsymbol{v}^T \boldsymbol{n} = 0$, where $\boldsymbol{n}$ denotes the surface normal. At a sufficiently large distance away from the airfoil, we prescribe a far field flow with a given Mach number of $Ma = 0.8$, pressure $p = 101\,325$ Pa and a temperature of 273.15 K. The uncertain angle of attack ϕ is uniformly distributed in the interval of [0.75, 1.75] degrees or in other words $\phi(\xi) = 1.25 + 0.5\xi$ with $\xi \sim U(-1, 1)$. The initial condition in the entire domain is equal to the far field boundary values and thus violates the Euler slip boundary condition at the airfoil. Consequently, we iterate in pseudo-time to correct the flow solution until the expectation value of the density fulfills the criterion (42) with $\varepsilon = 6 \cdot 10^{-6}$.

The used computational mesh (see Fig. 6c) consists of 22 361 triangular elements and resembles a circular domain of 40 m diameter, where the airfoil of 1 m length is located at the very center. The mesh is finely resolved close to the airfoil as we are only interested in effects close to the airfoil and becomes coarser the closer to the far field boundary. In order to be able to measure the quality of the obtained solutions with and without the One-Shot acceleration strategy, we compute a reference solution using stochastic-Collocation with 100 Gauss-Legendre quadrature points (see Fig. 6). We will show the L^2-error behavior of the discrete quantity $\boldsymbol{e}_\Delta = (\boldsymbol{e}_1, \cdots, \boldsymbol{e}_{N_x})^T$, where $\boldsymbol{e}_j$ is the cell average of the quantity $\boldsymbol{e}$ in spatial cell j. The discrete L^2 norm is denoted by

$$\|\boldsymbol{e}_\Delta\|_\Delta := \sqrt{\sum_{j=1}^{N_x} \Delta x_j \boldsymbol{e}_j^2} \,.$$

Given the SC reference solution $\boldsymbol{u}_\Delta$ and the moments of a compared numerical method $\hat{\boldsymbol{u}}_\Delta$, we investigate the relative error

$$\frac{\|\mathrm{E}[\boldsymbol{u}_\Delta] - \mathrm{E}[\mathcal{U}(\hat{\boldsymbol{u}}_\Delta)]\|_\Delta}{\|\mathrm{E}[\boldsymbol{u}_\Delta]\|_\Delta} \quad \text{and} \quad \frac{\|\mathrm{Var}[\boldsymbol{u}_\Delta] - \mathrm{Var}[\mathcal{U}(\hat{\boldsymbol{u}}_\Delta)]\|_\Delta}{\|\mathrm{Var}[\boldsymbol{u}_\Delta]\|_\Delta} \,.$$

As small fluctuations in the large cells of the coarse far field would dominate this error measure, we only compute the error inside a box of one meter height and 1.1 m length around the airfoil. Figure 7 shows the resulting error with respect to

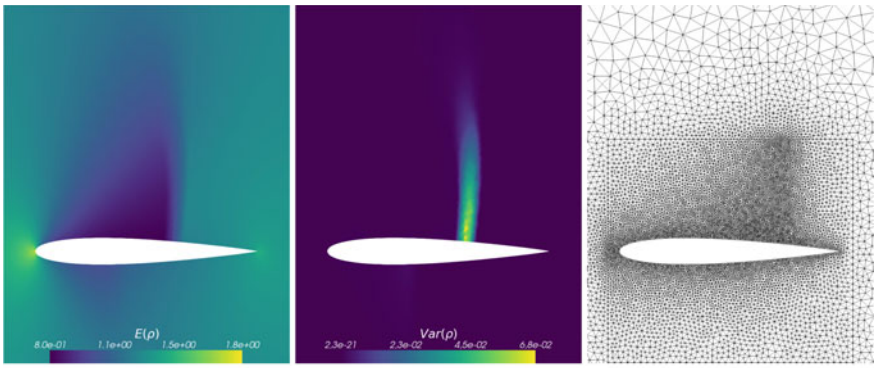


Fig. 6 Reference solution E[ρ] and Var[ρ] and the mesh close to the airfoil which is used in the computation of all presented methods.

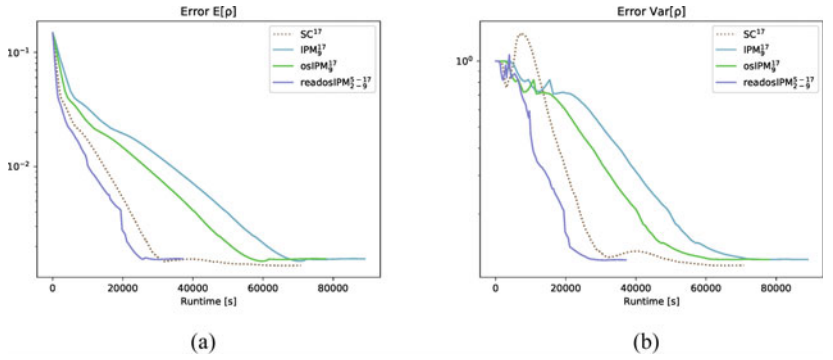


Fig. 7 Comparison of the relative L^2-error of the density for IPM, osIPM, readosIPM and SC. All IPM related methods are converged to a residual of $\varepsilon = 6 \cdot 10^{-6}$, whereas SC is converged to a residual of $\varepsilon = 1 \cdot 10^{-7}$. All computations are performed with 5 MPI threads.

the reference solution of IPM, osIPM and SC. The superscript in the figure denotes the number of used quadrature points, whereas the subscript denotes the moment order. We chose a total polynomial order of 9 for all IPM methods, meaning 10 moments are used in the computation and a Clenshaw-Curtis quadrature rule of order 4, resulting in 17 quadrature points. Based on the same quadrature set, all IPM solutions were also compared to a SC solution in order to get a better understanding of the methods convergence behavior and acceleration properties of osIPM. As it can be seen from the presented results, the osIPM yields the same error and almost identical convergence history as IPM, while being significantly faster. In comparison to SC however, the errors for the mean as well as the variance are comparably small, but the SC method reaches the given error level faster in terms of computational time. Only when the One-Shot approach is combined with adaptivity and refinement

Table 1 Moment and quadrature setup for the applied refinement levels.

Refinement level	0	1	2	3
Total degree of moments	1	2	3	4
Number of quadrature points	3	5	5	9

retardation, as it can be seen for the *readosIPM* plot in Fig. 7, the IPM method is yields even faster convergence rates than SC. For more information about refinement retardation and adaptivity, see [23]. Also note that the acceleration of the One-Shot approach becomes more dominant when looking at higher dimensional uncertainties, see [23].

6.4 Unsteady 2-D Euler Equations with Adaptivity

To investigate adaptivity, we again use the two dimensional Euler equations. The problem setting is similar to the transient Sod shock tube test case from above. The geometry describes a nozzle similar to a de Laval nozzle (see Fig. 8e), where the initial condition is set to a discontinuity positioned in the middle of the narrow part of the nozzle. The density in the left part is set to 1 kg m^{-3} and the energy is set to $\rho e = p/(\gamma - 1) = 2.5$ J m^{-3} with the pressure p equal to 1 Pa. For the right part of the domain the density is set to $\rho = 0.8$ kg m^{-3} and the pressure $p = 0.125$ Pa. The gas in both sides is at rest. For this testcase we inflict the initial condition with one uncertainty, i.e. the shock's position, which is now modeled as $x_{\text{shock}} \sim U(-0.5, 0.5)$. The used computational mesh consists of 76 696 triangular cells and is refined in the area of the shock and the nozzle opening towards the right side of the domain (see Fig. 8a). The applied boundary conditions are Euler slip conditions for the wall of the nozzle and Dirichlet conditions set to the initial condition for the left and right side of the mesh. The shown results in Fig. 8 resemble a time of 6 s.

As for the previous testcases, the reference solution was computed using stochastic-Collocation (see Figs. 8a, 8b) with a Gauss-Legendre quadrature with 50 quadrature points. As the previously mentioned parameter $\delta_{\pm}$ are user determined, these refinement/coarsening thresholds were set to $\delta_{-} = 1.5E-3$ and $\delta_{+} = 5E-4$ for the presented results in Figs. 8c, 8d. The resulting refinement levels are shown in Fig. 8c and the total order of used moments in combination with the associated quadrature points for each refinement level are given in Table 1. For the quadrature a tensorized Clenshaw-Curtis rule is used.

As for the previous test cases, we observe a good agreement between the IPM and the reference solution computed by SC. Due to the lower degree of moments the IPM solution again show a more step-like profile in the emerging shocks. As expected, the presented refinement levels are high in the regions around the shock and lower

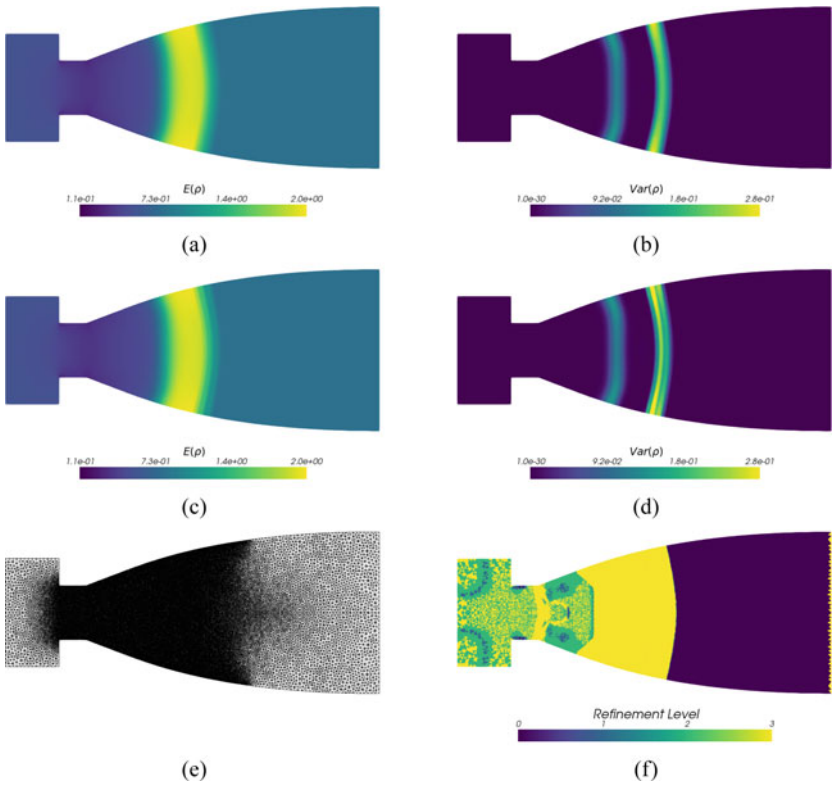


Fig. 8 Comparison for the mean and variance of the SC reference solution (**a**, **b**) and the adaptive IPM method (**c**, **d**) with the refinement levels in **f**. **e** shows the computational mesh. All computations are performed using 20 MPI threads.

order moments are started to be used further upstream where the flow becomes more and more constant as time progresses. Further upstream of the shock, the method even uses the lowest refinement level as the shock has not yet reached this part of the nozzle and thus the solution is still equal to the initial condition. All in all the results show that the method chooses high levels of refinement in areas where they are required by the complexity of the solution. Thus, the method is computationally much of efficient in the remainder of the domain.

Acknowledgment This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – FR 2841/6-1.

References

1. Alldredge, G., Hauck, C., Tits, A.: High-order entropy-based closures for linear transport in slab geometry II: a computational study of the optimization problem. SIAM J. Sci. Comput. **34**(4), B361–B391 (2012)
2. Alldredge, G., Hauck, C.D., O'Leary, D.P., Tits, A.L.: Adaptive change of basis in entropy-based moment closures for linear kinetic equations. J. Comput. Phys. **258**, 489–508 (2014)
3. Alldredge, G.W., Frank, M., Hauck, C.D.: A regularized entropy-based moment method for kinetic equations. SIAM J. Appl. Math. **79**(5), 1627–1653 (2019)
4. Bell, J.B., Dawson, C.N., Shubin, G.R.: An unsplit, higher order Godunov method for scalar conservation laws in multiple dimensions. J. Comput. Phys. **74**(1), 1–24 (1988)
5. Bijl, H., Lucor, D., Mishra, S., Schwab, C.: Uncertainty Quantification in Computational Fluid Dynamics, vol. 92. Springer (2013)
6. Caflisch, R.: Monte Carlo and Quasi-Monte Carlo methods. Acta Numerica (1998)
7. Colella, P.: Multidimensional upwind methods for hyperbolic conservation laws. J. Comput. Phys. **87**(1), 171–200 (1990)
8. Deshpande, S.: Kinetic theory based new upwind methods for inviscid compressible flows. In: 24th Aerospace Sciences Meeting, p. 275 (1986)
9. Kristopher Garrett, C., Hauck, C., Hill, J.: Optimization and large scale computation of an entropy-based moment closure. J. Computat. Phys. **302**, 573–590 (2015)
10. Ghanem, R., Higdon, D., Owhadi, H.: Handbook of Uncertainty Quantification. Springer (2017)
11. Giles, M.B.: Multilevel Monte Carlo methods. Acta Numerica (2018)
12. Guermond, J.-L., Nazarov, M., Popov, B., Yang, Y.: A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations. SIAM J. Numer. Anal. **52**(4), 2163–2182 (2014)
13. Harten, A., Lax, P.D., van Leer, B.: On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. SIAM Rev. **25**(1), 35–61 (1983)
14. Hauck, C.: High-order entropy-based closures for linear transport in slab geometry. Commun. Math. Sci. **9**, 187–205 (2011)
15. Hazra, S.B., Schulz, V., Brezillon, J., Gauger, N.R.: Aerodynamic shape optimization using simultaneous pseudo-timestepping. J. Comput. Phys. **204**(1), 46–64 (2005)
16. Holden, H., Risebro, N.H.: Front Tracking for Hyperbolic Conservation Laws, vol. 152. Springer (2015)
17. Kröker, I., Rohde, C.: Finite volume schemes for hyperbolic balance laws with multiplicative noise. Appl. Numer. Math. **62**(4), 441–456 (2012)
18. Kružkov, S.N.: First order quasilinear equations in several independent variables. Math. USSR-Sbornik **10**(2), 217 (1970)
19. Kusch, J., Alldredge, G.W., Frank, M.: Maximum-principle-satisfying second-order intrusive polynomial moment scheme. SMAI J. Comput. Math. **5**, 23–51 (2019)
20. Kusch, J., Frank, M.: Intrusive methods in uncertainty quantification and their connection to kinetic theory. Int. J. Adv. Eng. Sci. Appl. Math. **10**(1), 54–69 (2018)
21. Kusch, J., Frank, M.: An adaptive quadrature-based moment closure. Int. J. Adv. Eng. Sci. Appl. Math. **11**(3), 174–186 (2019)
22. Kusch, J., McClarren, R.G., Frank, M.: Filtered stochastic Galerkin methods for hyperbolic equations. J. Comput. Phys. **403**, 109073 (2019)
23. Kusch, J., Wolters, J., Frank, M.: Intrusive acceleration strategies for uncertainty quantification for hyperbolic systems of conservation laws (2019)
24. LeVeque, R.J.: Numerical Methods for Conservation Laws. Birkhäuser Verlag Basel (1992)
25. David Levermore, C.: Moment closure hierarchies for kinetic theories. J. Stat. Phys. **83**(5–6), 1021–1065 (1996)
26. Liu, X.-D.: A maximum principle satisfying modification of triangle based adapative stencils for the solution of scalar hyperbolic conservation laws. SIAM J. Numer. Anal. **30**(3), 701–716 (1993)

27. Le Maître, O.P., Knio, O.M.: Spectral Methods for Uncertainty Quantification. Springer (2010)
28. McClarren, R.G.: Uncertainty Quantification and Predictive Computational Science. Springer (2018)
29. Olbrant, E., Hauck, C.D., Frank, M.: A realizability-preserving discontinuous Galerkin method for the M1 model of radiative transfer. J. Comput. Phys. **231**(17), 5612–5639 (2012)
30. Persson, P.-O., Peraire, J.: Sub-cell shock capturing for discontinuous Galerkin methods. In: 44th AIAA Aerospace Sciences Meeting and Exhibit, p. 112 (2006)
31. Perthame, B.: Boltzmann type schemes for gas dynamics and the entropy property. SIAM J. Numer. Anal. **27**(6), 1405–1421 (1990)
32. Perthame, B.: Second-order Boltzmann schemes for compressible Euler equations in one and two space dimensions. SIAM J. Numer. Anal. **29**(1), 1–19 (1992)
33. Poëtte, G., Després, B., Lucor, D.: Uncertainty quantification for systems of conservation laws. J. Comput. Phys. **228**(7), 2443–2467 (2009)
34. Schlachter, L., Schneider, F.: A hyperbolicity-preserving stochastic Galerkin approximation for uncertain hyperbolic systems of equations. J. Comput. Phys. **375**, 80–98 (2018)
35. Smith, R.C.: Uncertainty Quantification: Theory, Implementation, and Applications. SIAM (2013)
36. Tadmor, E.: The numerical viscosity of entropy stable schemes for systems of conservation laws. I. Math. Comput. **49**(179), 91–103 (1987)
37. Tokareva, S., Schwab, C., Mishra, S.: High order SFV and mixed SDG/FV methods for the uncertainty quantification in multidimensional conservation laws. In: Abgrall, R., Beaugendre, H., Congedo, P.M., Dobrzynski, C., Perrier, V., Ricchiuto, M. (eds.) High Order Nonlinear Numerical Schemes for Evolutionary PDEs, pp. 109–133. Springer (2014)
38. Trefethen, L.N.: Cubature, approximation, and isotropy in the hypercube. SIAM Rev. **59** (2017)
39. Xiang, S., Bornemann, F.: On the convergence rates of gauss and Clenshaw-Curtis quadrature for functions of limited regularity. SIAM J. Numer. Anal. **50**(5), 2581–2587 (2012)
40. Xiu, D.: Numerical Methods for Stochastic Computations. Princeton University Press (2010)
41. Zhang, X., Shu, C.-W.: Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 467, pp. 2752–2776. The Royal Society (2011)

Well-Balanced Reconstruction Operator for Systems of Balance Laws: Numerical Implementation

I. Gómez-Bueno, M. J. Castro, and C. Parés

Abstract In some previous works, two of the authors introduced a strategy to develop high-order well-balanced numerical methods for 1d systems of balance laws. There, a strategy which allows us to modify any standard reconstruction operator in order to be well-balanced was also described. This strategy involves a nonlinear problem at every cell, at every time step, that consists in finding the stationary solution whose average is the given cell value. Our goal is to present a general efficient implementation that can be applied to any system of balance laws by interpreting these nonlinear problems as control problems that are rewritten in functional form. Newton's and descent methods are applied and compared. Applications to the Burgers' equation with a nonlinear source term and to the 1d shallow water model are finally shown.

1 Introduction

Let us consider a PDE system of the form:

$$U_t(x,t) + f(U(x,t))_x = S(U(x,t))H_x(x), \quad x \in \mathbb{R},\ t > 0, \tag{1}$$

where $U(x,t)$ takes values on an open convex set $\Omega \subset \mathbb{R}^N$, $f : \Omega \longrightarrow \mathbb{R}^N$ is the flux function, $S : \Omega \longrightarrow \mathbb{R}^N$, and $H : \mathbb{R} \longrightarrow \mathbb{R}$ is a continuous known function (possibly

I. Gómez-Bueno (✉) · M. J. Castro · C. Parés
University of Málaga, Málaga, Spain
e-mail: igomezbueno@uma.es

M. J. Castro
e-mail: mjcastro@uma.es

C. Parés
e-mail: pares@uma.es

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*,
SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2_3

the identity function). It is supposed that system (1) is strictly hyperbolic, that is, $D_f(U) = \dfrac{\partial f}{\partial U}(U)$ has N real different eigenvalues and eigenvectors.

Systems of the form (1) have non trivial stationary solutions that satisfy the ODE system:

$$f(U)_x = S(U)H_x. \tag{2}$$

A numerical method is said to be well-balanced if it solves exactly or with enhanced accuracy all the stationary solutions of the system or, at least, a relevant family of them. The use of methods with this property is of major importance when the waves generated for small perturbations of a steady state are going to be simulated: this is the case, for instance, for tsunami waves in the Ocean. Well-balanced methods have been studied by many authors: see [2] and its references for a recent review on this topic.

Recently, in [5] the following family of semidiscrete high-order well-balanced finite-volume methods for (1) has been discussed:

$$\frac{dU_i}{dt} = -\frac{1}{\Delta x}\left(F_{i+\frac{1}{2}}(t) - F_{i-\frac{1}{2}}(t)\right) + \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} S(P_i^t(x))H_x(x)\,dx, \tag{3}$$

where:

- $I_i = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$ are the computational cells, whose length Δx is supposed to be constant for simplicity;
- $U_i(t)$ is the approximation of the average of the exact solution at the ith cell at time t, that is,
$$U_i(t) \cong \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} U(x,t)\,dx;$$
- $P_i^t(x)$ is the approximation of the solution at the ith cell given by a high-order reconstruction operator from the sequence of cell averages $\{U_i(t)\}$:
$$P_i^t(x) = P_i(x; \{U_j(t)\}_{j\in\mathcal{S}_i});$$
where $\mathcal{S}_i$ denotes the set of indexes of the cells belonging to the stencil of the ith cell.
- $F_{i+\frac{1}{2}} = \mathbb{F}(U_{i+\frac{1}{2}}^{t,-}, U_{i+\frac{1}{2}}^{t,+})$, where $U_{i+\frac{1}{2}}^{t,\pm}$ are the reconstructed states at the intercells, i.e.
$$U_{i+\frac{1}{2}}^{t,-} = P_i^t(x_{i+\frac{1}{2}}), \quad U_{i+\frac{1}{2}}^{t,+} = P_{i+1}^t(x_{i+\frac{1}{2}}),$$
and $\mathbb{F}$ is a consistent first order numerical flux.

It can be then easily shown that, if the reconstruction operator is well-balanced for a stationary solution U of (1) then the numerical method is also well-balanced for U according to the following definitions:

Definition. Given a stationary solution U of (1):

- The numerical method (3) is said to be well-balanced for U if the vector of cell averages of U is an equilibrium of the ODE system (3).
- The reconstruction operator is said to be well-balanced for U if

$$P_i(x) = U(x), \quad \forall x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}], \ \forall i, \tag{4}$$

where P_i is the approximation of U obtained by applying the reconstruction operator to the vector of cell averages of U. □

The following strategy to design a well-balanced reconstruction operator P_i on the basis of a standard operator Q_i was introduced in [1]: given a family of cell values $\{U_i\}$, at every cell $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$:

1. Look for the stationary solution $U_i^*(x)$ such that:

$$\frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} U_i^*(x)\, dx = U_i. \tag{5}$$

2. Apply the reconstruction operator to the cell values $\{\widetilde{U}_j\}_{j \in S_i}$ given by

$$\widetilde{U}_j = U_j - \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} U_i^*(x)\, dx,$$

to obtain:

$$Q_i(x) = Q_i(x; \{\widetilde{U}_j\}_{j \in \mathcal{S}_i}).$$

3. Define

$$P_i(x) = U_i^*(x) + Q_i(x). \tag{6}$$

It can be then easily shown that the reconstruction operator P_i in (6) is well-balanced for every stationary solution provided that the reconstruction operator Q_i is exact for the null function. Moreover, P_i is conservative, i.e.

$$\frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} P_i(x)\, dx = U_i, \quad \forall i, \tag{7}$$

provided that Q_i is conservative, and it is also high-order accurate provided that the stationary solutions are smooth.

The main difficulty when this strategy is applied comes from the first step of the well-balanced reconstruction operator: a nonlinear problem of the form (5) has to be solved at every time step. Since the stationary solutions of (1) are the solutions of the ODE system (2), problem (5) is equivalent to find the solution of an ODE system with prescribed average in an interval. In some cases, the explicit form of the general

solution of the ODE is known and (5) can be solved by hand or by using standard iterative methods for nonlinear problems: see [5] for examples.

Our goal is to describe a general methodology to solve numerically problems of the form (5) and to apply it to the implementation of well-balanced reconstruction operators for general systems of balance laws whether or not the analytical expression of the stationary solutions is known.

The organization of this work is as follows: in Sect. 2 problem (5) is interpreted as a control problem. It is first written in functional form; then, the gradient of the functional is computed using the adjoint equation. Newton's and descent methods can be applied to solve numerically the problem: this is done in Sect. 3. A comparison between both methods is shown in Sect. 4, where different strategies for the choice of the descent step have been also tried. In practice, the state and the adjoint equations, the cell-averages, and the integral appearing at the source terms are computed numerically: two numerical tests are shown in Sect. 5 in order to check the accuracy and the well-balancedness of the methods. A scalar balance law and the shallow water model are considered. Finally, some conclusions are drawn.

2 Control Problem

In the first stage of the well-balanced reconstruction procedure one has to find a solution of the ODE problem

$$f(U)_x = S(U)H_x, \tag{8}$$

such that its average in the cell $[x_{i-1/2}, x_{i+1/2}]$ is the given cell value U_i. For stationary solutions such that $D_f(U(x))$ is regular for every x (i.e. no sonic state is reached), solving (8) is equivalent to solve the ODE system in normal form:

$$U_x = G(x, U), \tag{9}$$

where G is the function $G : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}^N$ defined by

$$G(U, x) = D_f(U)^{-1} S(U) H_x. \tag{10}$$

In this article we focus on the preservation of stationary solutions satisfying (9): the methods introduced here preserve supersonic or subsonic stationary solutions, but no transcritical ones. As it will be seen, in the algorithm developed here to compute the first stage of the well-balanced operator, Cauchy problems associated to the ODE system (9) will be numerically solved, what can be done with any standard ODE solver. To adapt the algorithm to the preservation of transcritical stationary solutions, Cauchy problems associated to (9) should be numerically computed, what can be difficult when the initial condition is a sonic state or if it is close to it. In this case the system may have no solution or to have more than one. The strategy followed here to

deal with this difficulty consists on applying the standard reconstruction procedure whenever a sonic state is detected in the stencil. This simple modification allows us to overcome this difficulty and the algorithm is only modified in the stencils where a sonic point is detected. In Sect. 5.2.2 it will be seen that, in the case of the shallow water equations, this strategy allows one to correctly handle with transonic regimes and even to accurately preserve transonic stationary solutions. Nevertheless, to extend the algorithm so that these solutions are genuinely preserved is a challenging problem that is out of the scope of this work and will be faced in a forthcoming work.

Notice that, even if only stationary solutions satisfying (9) are sought, the problem consisting in finding a solution of this ODE system with given average may have no solution. Observe that if (5) has no solution at the ith cell then U_i cannot be the average of any stationary solution. Therefore, at this cell the standard reconstruction operator is applied, i.e. $U_i^* \equiv 0$ is chosen in the first step.

Let us assume thus that no sonic points have been detected in the stencil $\mathcal{S}_i$. Then, the nonlinear problem (5) to be solved at every cell can be then formulated as follows:

Find $U_{i-1/2} \in \Omega$ such that

$$\mathcal{F}(U_{i-1/2}) = U_i, \tag{11}$$

where $\mathcal{F} : \Omega \mapsto \mathbb{R}^N$ is given by

$$\mathcal{F}(U_0) = \frac{1}{\Delta x} \int_0^{\Delta x} V_i(x, U_0)\, dx, \tag{12}$$

where $V_i(x, U_0)$ denotes the solution of the Cauchy problem

$$\begin{cases} V_x = G(V, x + x_{i-1/2}), \\ V(0) = U_0. \end{cases} \tag{13}$$

Notice that Cauchy problem (13) is equivalent to solve (9) with initial condition

$$U(x_{i-1/2}) = U_{i-1/2}.$$

Once (11) has been solved, the sought stationary solution is

$$U_i^*(x) = V(x - x_{i-1/2}, U_{i-1/2}).$$

This problem can be interpreted as a control one, and the adjoint technique can be used to compute the gradient of $\mathcal{F}$, what gives:

$$D\mathcal{F}(U_0) = \frac{1}{\Delta x} \Lambda(0)^T,$$

where Λ denotes the matrix whose columns are the so-called adjoint variables $\boldsymbol{\lambda}_1(x), \ldots, \boldsymbol{\lambda}_N(x)$ that solve the following Cauchy problems:

$$\begin{cases} \dfrac{d\boldsymbol{\lambda}_j}{dx}(x) = -\mathbf{e}_j - \nabla_U G(U, x + x_{i+1/2})^T \cdot \boldsymbol{\lambda}_j, \\ \\ \boldsymbol{\lambda}_j(\Delta x) = 0, \end{cases} \tag{14}$$

where we denote by $\mathbf{e}_j$ the jth vector of the canonical basis and

$$\nabla_U G(U, x) = \begin{bmatrix} \dfrac{\partial G_1}{\partial u_1}(U, x) & \dots & \dfrac{\partial G_1}{\partial u_N}(U, x) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial G_N}{\partial u_1}(U, x) & \dots & \dfrac{\partial G_N}{\partial u_N}(U, x) \end{bmatrix}; \tag{15}$$

(see [6] for details).

First and second order methods can be implemented in an easier way if the midpoint rule is used to approach the cell averages:

$$\frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} U(x)\, dx \cong U(x_i).$$

In effect, in this case the first step in the reconstruction procedure reduces to look for the stationary solution U_i^* such that:

$$U_i^*(x_i) = U_i. \tag{16}$$

There is no need thus to solve nonlinear problems of the form (11): it is enough to solve standard Cauchy problems to compute U_i^*. Therefore, in what follows we focus on methods of order greater than two. In particular, the third order CWENO reconstruction (see [3, 9]) will be considered in the numerical experiments. The state and the adjoint differential equations will be numerically computed using the standard RK4 method.

3 Numerical Algorithms

Let us consider for simplicity that $x_{i-1/2} = 0$ and denote by $W \in \Omega$ the cell value, so that the problem to solve is:

Find $U_0 \in \Omega$ such that

$$\mathcal{F}(U_0) = W. \tag{17}$$

Since problems of this type have to be solved at every intercell at every time step, it is crucial to choose an efficient numerical method. Two different strategies are considered here:

3.1 Newton's Method

A sensible choice for the initial guess U_0^0 is W: if Δx is small, the average of the solution of the Cauchy problem is expected to be close to the initial condition. The algorithm is then as follows:

Algorithm. Newton's method

- $U_0^0 = W$;
- For k = 0,1,2…
  - Compute the solution U_k of (13) with initial condition U_0^k in the interval $[0, \Delta x]$.
  - For $j = 1, \ldots, N$ compute the solution $\boldsymbol{\lambda}_j$ of (14) with $U = U_k$ in the interval $[0, \Delta x]$.
  - Compute V_k by solving the linear system:

$$\Lambda(0)^T V_k = \Delta x(\mathcal{F}(U_0^k) - W).$$

  - Update U_0^k:

$$U_0^{k+1} = U_0^k - V_k.$$

At every iteration of the method $N + 1$ Cauchy problems and a $N \times N$ linear system have to be solved. The computational cost can be reduced by using the modified Newton's method in which the matrix $\Lambda(0)$ is only updated every K iterations, where K is a fixed integer.

3.2 Descent Methods

An alternative approach to solve (5) consists in solving the minimization problem:

$$\min_{U_0 \in \mathbb{R}^N} J(U_0) \tag{18}$$

with:

$$J(U_0) = \|\mathcal{F}(U_0) - W\|^2.$$

If the euclidean norm is chosen, a simple computation shows that:

$$\nabla J(U_0) = \frac{2}{\Delta x} \Lambda(0)^T \cdot \left(\frac{1}{\Delta x} \int_0^{\Delta x} (U(x, U_0) - W)\, dx \right). \tag{19}$$

Gradient method writes thus as follows:

Algorithm. Gradient method

- $U_0^0 = W$.
- For $k = 0, 1, \ldots$:
  - Compute the solution U_k of (13) with initial condition U_0^k in the interval $[0, \Delta x]$.
  - For $j = 1, \ldots, N$ compute the solution $\boldsymbol{\lambda}_j$ of (14) with $U = U_k$ in the interval $[0, \Delta x]$.
  - Compute

$$\nabla J(U_k) = \frac{1}{\Delta x} 2\, \Lambda(0)^T \cdot \left(\frac{1}{\Delta x} \int_0^{\Delta x} (U_k - W)\, dx \right).$$

  - Update U_0^k:

$$U_0^{k+1} = U_0^k - \rho_k d_k,$$

  where ρ_k is the step and $d_k = \nabla J(U_k)$.

Conjugate gradient methods are also considered with the descent directions:

$$d_k = \begin{cases} \nabla J(U_k) & if\ k = 0, \\ \nabla J(U_k) + \dfrac{\nabla J(U_k) \cdot (\nabla J(U_k) - \nabla J(U_{k-1}))}{\|\nabla J(U_{k-1})\|^2}\, d_{k-1} & if\ k \geq 1. \end{cases}$$

At every iteration of the gradient or the conjugate gradient methods, at least $N + 1$ Cauchy problems have to be solved, but a new Cauchy problem has to be solved in every evaluation of the cost function in the search of the step ρ_k.

Search for the Optimal Step

Once the descent direction has been computed, the step ρ_k has to be chosen. Five different options of stepsize selection are discussed.

β **Stepsize Method. First Version** At the kth iteration, when steps 1 and 2 of the descent algorithm have already been computed, use the following strategy to select the stepsize:

Algorithm β **Stepsize Method. First Version.**

- Set $\rho_k^0 = \rho_{k-1}$. (If $k = 1$, ρ_{-1} is arbitrarily chosen).
- Set $V_k^0 = J(U_{k-1})$
- Set $j = 1$:

- Set $V_k^j = J(U_{k-1} - \rho_k^{j-1} d_{k-1})$.
- If $V_k^j < V_k^{j-1}$, choose $\rho_k^j = \beta \rho_k^{j-1}$. In other case, choose $\rho_k^j = \dfrac{\rho_k^{j-1}}{\beta}$.
- If $j \geq 2$ y $\rho_k^j = \rho_k^{j-2}$, stop.

- Set $\rho_k = \rho_k^j$.
- $j = j + 1$. □

Here β is a parameter to be chosen in the interval $(0, 1)$.

β Stepsize Method. Second Version. At the kth iteration, when steps 1 and 2 of the descent algorithm have already been computed, use the following strategy to select the stepsize:

***Algorithm* β stepsize method. Second version**

- Set $V_{0,k}^d = J(U_{k-1})$, $d_{k-1} = \nabla J(U_{k-1})$.
- $\rho_{0,k}^d = \beta \rho_{k-1}$. (If $k = 1$, ρ_{-1} is arbitrarily chosen).
- Set $V_{1,k}^d = J(U_{k-1} - \rho_{0,k}^d d_{k-1})$.
- Set $j = 1$.
- While $V_{j,k}^d < V_{j-1,k}^d$ do:
  - Set $\rho_{j,k}^d = \beta\, \rho_{j-1,k}^d$ and $V_{j+1,k}^d = J(U_{k-1} - \rho_{j,k}^d d_{k-1})$.
  - $j = j + 1$.
- Set $V_{0,k}^h = J(U_{k-1})$.
- $\rho_{0,k}^h = \dfrac{1}{\beta} \rho_{k-1}$.
- Set $V_{1,k}^h = J(U_{k-1} - \rho_{0,k}^h d_{k-1})$.
- Set $j = 1$.
- While $V_{j,k}^h < V_{j-1,k}^h$ do:
  - Set $\rho_{j,k}^h = \dfrac{1}{\beta}\, \rho_{j-1,k}^h$ and $V_{j+1,k}^h = J(U_{k-1} - \rho_{j,k}^h d_{k-1})$.
  - $j = j + 1$.
- If $V_{j,k}^d < V_{j,k}^h$, set $\rho_k = \rho_{j,k}^d$. Else, $\rho_k = \rho_{j,k}^h$. □

β is again a parameter to be chosen in the interval $(0, 1)$.

Armijo Rule. The following algorithm is based on the Armijo Rule. It requires two parameters: $\mu \in [0.01, 0.3]$ and $\beta \in [0.1, 0.8]$. Let us suppose that $\bar{U}$ is the approximation of the solution at the previous iteration and $\bar{d}$ is the descent direction. Let us consider the function $\mathbf{J}(\rho) = J(\bar{U} - \rho \bar{d})$. Then the first order approximation of $\mathbf{J}(\rho)$ at $\rho = 0$ is given by $\mathbf{J}(0) - \rho\, \mathbf{J}'(0)$. Define $\hat{\mathbf{J}}(\rho) = \mathbf{J}(0) - \mu\, \rho\, \mathbf{J}'(0)$. A stepsize $\bar{\rho}$ is considered acceptable by the Armijo Rule if $\mathbf{J}(\bar{\rho}) \leq \hat{\mathbf{J}}(\bar{\rho})$.

When the steps 1, 2 and 3 of the general algorithm described above have been computed at the kth iteration, define the functions:

$$\mathbf{J}(\rho_k) = J(U_k - \rho_k d_k), \quad \hat{\mathbf{J}}(\rho_k) = \mathbf{J}(0) - \mu\, \rho_k \mathbf{J}'(0) = J(U_k) - \mu\, \rho_k d_k \cdot d_k.$$

The stepsize ρ_{k+1} is chosen using the next rule: if $\mathbf{J}(\rho_k) > \hat{\mathbf{J}}(\rho_k)$, then $\rho_{k+1} = \beta\rho_k$, and otherwise take $\rho_{k+1} = \rho_k$.

This algorithm can be improved by applying the Armijo Rule as many times as possible at every iteration, provided that the objective function decreases as in the second version of the β stepsize method.

Wolfe Conditions

The choice of the step is based on Wolfe conditions: see [8]. It requires two parameters: $m_1 \in [0.01, 0.3]$ and $m_2 \in [0.5, 1]$. At the kth iteration, once steps 1, 2 and 3 of the general algorithm have been computed, follow the following algorithm:

***Algorithm* Wolfe Conditions**

a. Set $\rho_s^k = 0$ and $\rho_b^k = 0$.

b. Compute the functions:

$$\mathbf{J}(\rho_k) = J(U_k - \rho_k d_k),$$

$$\hat{\mathbf{J}}(\rho_k) = J(U_k) - \mu\, \rho_k\, (d_k \cdot d_k)\,.$$

– If $\mathbf{J}(\rho_k) \leq \hat{\mathbf{J}}(\rho_k)$ and $\mathbf{J}'(\rho_k) \geq m_2\, \mathbf{J}'(0)$, take $\rho_{k+1} = \rho_k$ and stop the algorithm.
– If $\mathbf{J}(\rho_k) > \hat{\mathbf{J}}(\rho_k)$, set $\rho_b^k = \rho_k$ and go to the step **c**.
– If $\mathbf{J}(\rho_k) \leq \hat{\mathbf{J}}(\rho_k)$ and $\mathbf{J}'(\rho_k) < m_2\, \mathbf{J}'(0)$, set $\rho_s^k = \rho_k$ and go to **c**.

c. Use the following rule to choose ρ_{k+1} :

– If $\rho_b^k = 0$, take $\rho_{k+1} = a\rho_k$, where $a > 1$.
– Else, take $\rho_{k+1} = \dfrac{\rho_s^k + \rho_b^k}{2}$. □

Fixed Stepsize

The first step ρ_0 is computed using any of the previous algorithms and then

$$\rho_k = \rho_0, \quad \forall k.$$

3.3 Numerical Integration

In practice, a quadrature rule in $[0,\, \Delta x]$ is used to compute the averages appearing in the definition of $\mathcal{F}$ (12), in the expression of the gradients (19), or in the computation of W (if it is the average of a known function):

$$\int_0^{\Delta x} g(x)\, dx \cong \Delta x \sum_{l=0}^{M} \alpha_l\, g(x_l),$$

and the Cauchy problems to compute U_k and $\boldsymbol{\lambda}_j$ are solved with a numerical method for ODE using a mesh of the interval $[0,\, \Delta x]$ whose maximum step will be denoted

by h. This mesh will be chosen so that all the quadrature points x_l are nodes. The order of the method and the size of h will be chosen so that errors are close to machine precision.

Therefore, in practice the following discrete problems have to be solved :

Find U_0 such that

$$\mathcal{F}_h(U_0) := \sum_{l=0}^{M} \alpha_l U_{h,l} = \widetilde{W},$$

where $U_{h,l}$ represents the numerical approximation of $U(x, U_0)$ at the quadrature point x_l given by the numerical method chosen to solve the ODE and $\widetilde{W}$ the approximation of W obtained with que quadrature formula.

4 A Numerical Test for the Control Problem

The Newton's and descent methods with different strategies for the choice of the descent steps have been tested and compared in order to check their efficiencies.

We will discuss the following scalar problem: find $u_0 \in \mathbb{R}$ such that

$$\mathcal{F}_{\Delta x}(u_0) = w, \tag{20}$$

where

$$\mathcal{F}_{\Delta x}(u_0) = \frac{1}{\Delta x} \int_0^{\Delta x} u(x, u_0)\, dx,$$

$w = 1$, and $u(x, u_0)$ is the solution of:

$$\begin{cases} u_x = \dfrac{\sin(u)}{u}, \\ u(0) = u_0. \end{cases} \tag{21}$$

The two-points Gauss quadrature rule is considered to approximate the integrals and the fourth order Runge-Kutta method is applied to solve the Cauchy problems using the mesh consisting of the extremes of the interval $[0, \Delta x]$ and the two quadrature points.

In order to check the efficiency of the Newton's and descent methods we compare the number of iterations and the CPU times required for solving problem (21) a big number of times (say 10000 times) with decreasing values of Δx and different error tolerances ε (see Tables 1 and 2). The values of Δx considered are in the range of those used in applications to the numerical solution of systems of balance laws: remember that problems (11) are solved at every computational cell. For these small values of Δx, the number of iterations of the gradient methods is independent in this case of the strategy followed to select the stepsizes.

Table 1 Number of iterations for different intervals $[0, \Delta x]$

Δx	Tolerance $\varepsilon = 10^{-8}$		Tolerance $\varepsilon = 10^{-12}$	
	Newton's method	Gradient methods	Newton's method	Gradient methods
1	2	6	3	10
0.5	2	4	2	7
0.1	1	2	1	4

Table 2 CPU times in milliseconds for different intervals $[0, \Delta x]$

Δx	Newton's method	Gradient method				
		$\beta = 0.5$ 1V.	$\beta = 0.5$ 2V.	Armijo 1V.	Armijo 2V.	Wolfe
Tolerance $\varepsilon = 10^{-8}$						
1	16	47	94	31	31	45
0.5	10	31	63	16	16	31
0.1	2	19	36	11	11	18
Tolerance $\varepsilon = 10^{-12}$						
1	31	210	266	125	125	188
0.5	10	142	168	78	78	131
0.1	2	105	136	61	61	79

The value $\rho_0 = 0.5$ is considered as the initial stepsize for the descent methods. We denote by $\beta = 0.5$ 1V the first version of the β stepsize algorithm taking $\beta = 0.5$, and $\beta = 0.5$ 2V is used for the second version. For both versions of the strategy based on the Armijo rule, $\mu = 0.1$ and $\beta = 0.5$ have been used as parameters and $m_1 = 0.1$, $m_2 = 0.6$ and $a = 2.0$ are taken when the Wolfe conditions are considered.

In both versions of the β stepsize method, $\beta \in (0, 1)$ is a parameter to be chosen. Up to now, $\beta = 0.5$ has been chosen. In order to study the influence of this parameter in the numerical simulations, the problem test has been solved for different values of β, taking $\varepsilon = 10^{-12}$ as tolerance and $\Delta x = 1$ (see Table 3).

Notice that, in both versions of the algorithm, the number of iterations decreases as β tends to 1 and the computational cost increases as β tends to 0 or to 1. In view of the results, the first version of the β stepsize method with $\beta = 0.9$ seems to be the best choice for this problem test.

As it can be observed, in spite of the bigger number of Cauchy problems to be solved at every iteration, Newton's method perform better than gradient methods both in number of iterations and in CPU time. Moreover, the differences increase as Δx and the tolerance decrease.

Furthermore, the fixed stepsize strategy combined with all the strategies considered to select the step has been also tested. Despite the fact that the CPU times obtained when applying this strategy are smaller than the ones shown in Table 2, Newton's method is still cheaper in every case. The conjugate gradient method,

Table 3 Number of iterations and computational time for different values of β

β stepsize method				
β	First version		Second version	
	Iterations	CPU time	Iterations	CPU time
0.1	10	212	10	266
0.2	10	210	10	267
0.3	10	211	10	266
0.4	10	210	10	266
0.5	10	210	10	266
0.6	10	212	9	245
0.7	6	152	6	234
0.8	5	139	5	215
0.9	4	125	4	190
0.99	4	290	3	320

which has been also applied, increases the computational cost in relation to the gradient methods. Therefore, we conclude that in our case, the Newton's method is the most efficient strategy in order to solve the control problem.

In order to confirm this statement, a similar study has been performed in the numerical examples considered in the next Section. The conclusion has been the same in all cases: for the typical values of Δx used in the applications, Newton's method is always the most efficient in terms of the computational cost.

5 Numerical Experiments

In order to implement the high-order well-balanced numerical methods introduced in Section 1 we consider:

- Rusanov numerical flux;
- the third order CWENO reconstructions (see [3, 9]);
- the third order TVD Runge-Kutta for solving the ODE system (3): see [7];
- the Gauss two points quadrature rule;
- the standard fourth order Runge-Kutta method for solving the state and the adjoint ODE problems related to the control problem at every cell $[x_{i-1/2}, x_{i+1/2}]$. The submesh considered in the cell consists of three $(N_p + 1)$-point uniform partitions of the subintervals

$$[x_{i-1/2}, x_0^i], [x_0^i, x_1^i], [x_1^i, x_{i+1/2}],$$

where x_j^i, $j = 0, 1$ are the quadrature points. The total number of points of the submesh is thus of $3N_p + 1$

When the initial condition is a stationary solution U^* in an interval $[a, b]$, we approximate its cell averages either by applying the quadrature formula to the exact solution (when it is available) or by

$$U^*_{h,i} = \sum_{l=0}^{M} \alpha^i_l U^{*,i}_{h,l}$$

where $U^{*,i}_{h,j}$ are the approximations at the quadrature points obtained using RK4 to approximate (2) with initial condition

$$U(a) = U^*(a).$$

Observe that the only information about the particular problem required by the numerical method is f, S, H, G, ∇G (see (1), (10), (15)) what leads to very general algorithms.

The following symbols will be used in this section to denote the different methods considered:

- SM3: numerical method of third order based on the Rusanov flux and the standard reconstruction operators.
- NWBM3: numerical method of third order based on the Rusanov flux and the well-balanced reconstruction operators in which problems (5) are solved numerically using the Newton's method.

5.1 Burgers Equation with a Nonlinear Source Term

We consider Burgers equation with a non-linear source term:

$$\begin{cases} u_t + \left(\dfrac{u^2}{2}\right)_x = \sin(u), \quad x \in \mathbb{R},\ t > 0, \\ u(x, 0) = u_0(x). \end{cases} \tag{22}$$

This problem is the particular case of (1) corresponding to:

$$U = u, \quad f(U) = \frac{u^2}{2}, \quad S(U) = \sin(u), \quad H(x) = x.$$

The ODE satisfied by the stationary solutions is

$$\frac{du}{dx} = \frac{\sin(u)}{u}. \tag{23}$$

Therefore:

$$G(x,U)=\frac{\sin(u)}{u},\quad \partial_U G(x,U)=\frac{u\cos(u)-\sin(u)}{u^2}.$$

In this case, the stationary solutions satisfy the EDO studied in the test problem introduced in Sect. 4. These stationary solutions cannot be expressed in terms of elementary functions so that (5) has to be numerically solved.

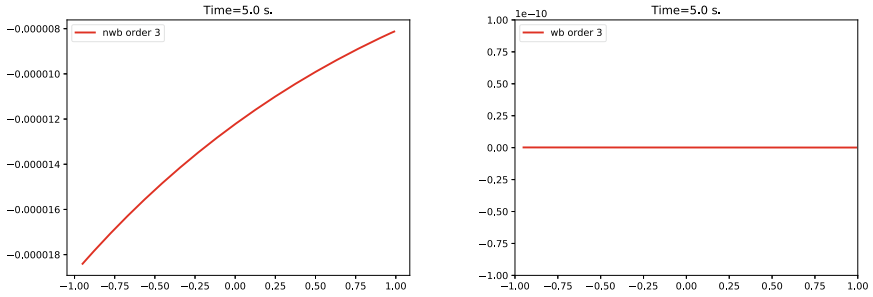
We consider $x\in[-1,1]$, $t\in[0,5]$ and $CFL=0.9$. The initial condition is the solution of the Cauchy problem consisting of (23) with initial condition

$$u(-1)=2,$$

which is a stationary solution of the problem. This solution is approximated using the RK4 method and $N_p=1$ is considered.

$u(-1,t)=2$ is imposed at $x=-1$ and free boundary conditions are considered at $x=1$.

Figure 1 shows the differences between the stationary solution and the numerical results obtained with SM3 (left) and NWBM3 (right). Table 4 show the L^1-errors and the empirical order of convergence of SM3. Notice that the non well-balanced methods perturb the stationary solution.



(a) SM3 (b) NWBM3

Fig. 1 Test 5.1. Differences between the stationary solution and the numerical solutions at $t=5s$. Number of cells: 100

Table 4 Test 5.1. Errors in L^1 norm and convergence rates for SM3 and errors in L^1 norm and convergence rates for NWBM3

Cells	SM3		NWBM3
	Error	Order	Error
100	7.66E−5	–	2.54E−13
200	9.62E−8	10.506	3.60E−14
400	1.21E−10	7.254	2.12E−14
800	1.51E−11	2.922	9.11E−14

The maximum number of iterations required to solve the nonlinear problem (5) applying Newton's method is two and it converges in only one iteration for meshes with 200 cells or more. As expected, the well-balanced modification increases the computational effort. The computational cost required for solving the problem with 100 cells is 50 ms for SM3 and 760 ms if NWBM3 is applied. If the number of cells considered is 200, the CPU time for SM3 is 190 ms, whereas for NWBM3 is 2220 ms. In any case, this extra computational cost is lower than the one that would be required to lead the discretization errors to close to zero machine by refining the mesh or increasing the order of non-well-balanced methods.

5.2 Shallow Water Equations

Let us consider the shallow water model, which is the particular case of (1) corresponding to the choices $N = 2$,

$$U = \begin{pmatrix} h \\ q \end{pmatrix}, \quad f(U) = \begin{pmatrix} q \\ \dfrac{q^2}{h} + \dfrac{g}{2}h^2 \end{pmatrix}, \quad S(U) = \begin{pmatrix} 0 \\ gh \end{pmatrix}.$$

The variable x makes reference to the axis of the channel and t is the time; $q(x, t)$ and $h(x, t)$ are the discharge and the thickness, respectively; g is the gravity and $H(x)$ is the depth function measured from a fixed reference level. We denote by $u = q/h$ the depth-averaged velocity and $c = \sqrt{gh}$.

The eigenvalues of the Jacobian matrix $D_f(U)$ of the flux function $f(U)$ are the following:

$$r_1 = u - \sqrt{c}, \quad r_2 = u + \sqrt{c}.$$

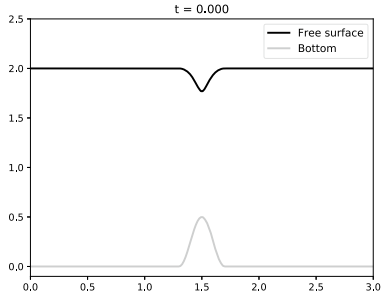
The Froude number, given by $Fr(U) = |u|/c$ indicates the flow regime: subcritical ($Fr < 1$), critical ($Fr = 1$) or supercritical ($Fr > 1$).

If $Fr(U) \neq 1$ the system of ODE satisfied by the stationary solutions is

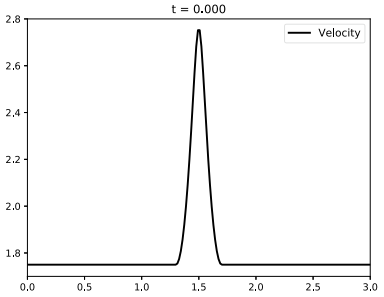
$$\begin{cases} q_x = 0, \\ h_x = \dfrac{gh H_x}{-u^2 + gh}. \end{cases} \tag{24}$$

5.2.1 A Subcritical Stationary Solution

Let us consider a test case taken from [4]: $x \in [0, 3], t \in [0, 5]$, and the depth function is given by:



(a) Initial condition. Free surface and bottom.



(b) Initial condition. Velocity.

Fig. 2 Test 5.2.1. Initial condition: a subcritical stationary solution computed with RK4

$$H(x) = \begin{cases} -0.25(1 + \cos(5\pi(x + 0.5))) \text{ if } 1.3 \leq x \leq 1.7, \\ \\ 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{25}$$

As initial condition, we consider the subcritical stationary solution of (24) with initial conditions $h(0) = 2,\ q(0) = 3.5$, (see Fig. 2).

As boundary conditions, $q(0, t) = 3.5$ is set upstream, while the water height is imposed to be $h(3, t) = 2.0$ downstream. The CFL parameter is set again to 0.9. The conclusions are similar to the previous test case: Fig. 4 shows the differences between the stationary solution and the numerical results obtained with SM3 (up) and NWBM3 with $N_p = 3$ (down). Table 5 shows the L^1-errors and the empirical order of convergence for SM3 and NWBM3 with $N_p = 1$ and $N_p = 3$. Errors in NWBM3 are due to the numerical approximation of the stationary solution with RK4 and thus the empirical order of convergence is 4. In any case they are significantly lower than those corresponding to SM3.

Concerning the computational cost, we have checked the effect of using Newton's method or its modification in which the adjoint variable is recomputed every K iterations. Since, in this case, the maximum number of iterations of Newton's method throughout the computations is 6, we have compared the computational effort for values of K ranging from 1 (the adjoint variable is recomputed at every iteration) to 6 (it is only computed once at the beginning in all cases): Fig. 3 shows the CPU times for the third order method. As it can be seen, the best option is to solve the adjoint problem only once at the beginning.

Again, the well-balanced procedure increases the computational effort. The computational cost required for solving the problem with 100 cells is 300 ms for SM3 and 2010 ms if NWBM3 with $N_p = 1$ is applied and with 200 cells, the CPU time for SM3 is 1020 ms, whereas for NWBM3 with $N_p = 1$ is 7690 ms. The computational cost increases linearly with N_p.

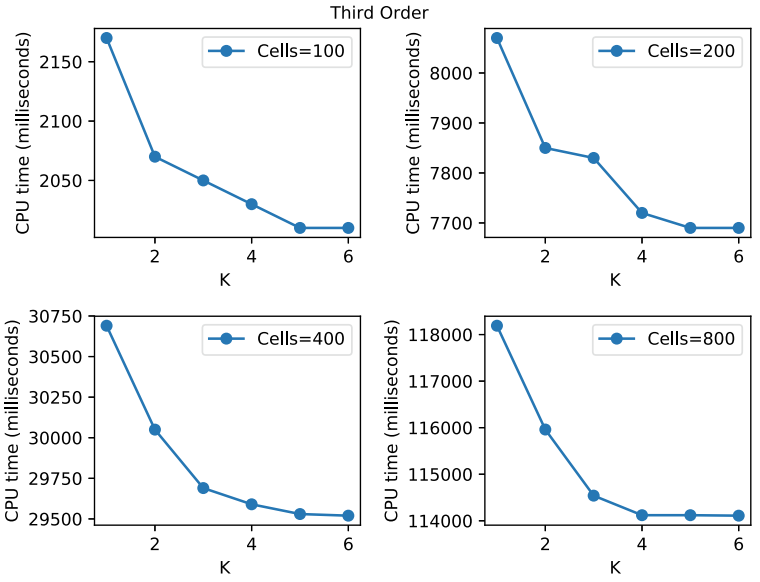
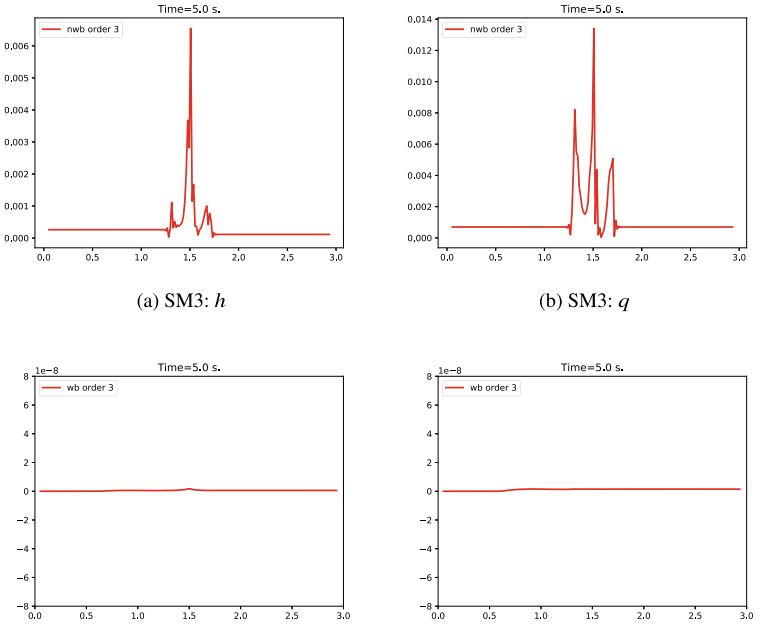


Fig. 3 Test 5.2.1. CPU times corresponding to NWBM3 with different number of cells and different values of K



(a) SM3: h (b) SM3: q

(c) NWBM3 (with $N_p = 3$): h (d) NWBM3 (with $N_p = 3$): q

Fig. 4 Test 5.2.1. Differences between the stationary solution and the numerical solutions at $t = 5s$. Number of cells: 200

Table 5 Test 5.2.1. Errors in L^1 norm and convergence rates for SM3 and NWBM3

Cells	SM3		NWBM3			
			$N_p = 1$		$N_p = 3$	
	Error	Order	Error	Order	Error	Order
h						
100	5.98E−3	–	7.45E−6	–	3.75E−8	–
200	9.16E−4	2.707	1.93E−7	5.271	1.40E−9	4.743
400	1.21E−4	2.920	6.63E−9	4.863	6.50E−11	4.429
800	1.60E−5	2.919	2.42E−10	4.776	3.23E−12	4.331
q						
100	2.12E−2	–	1.83E−5	-	9.16E−8	–
200	3.23E−3	2.714	4.70E−7	5.283	3.39E−9	4.756
400	4.26E−4	2.923	1.62E−8	4.859	1.58E−10	4.423
800	5.47E−5	2.961	5.94E−10	4.769	8.53E−12	4.051

5.2.2 A Transcritical Solution

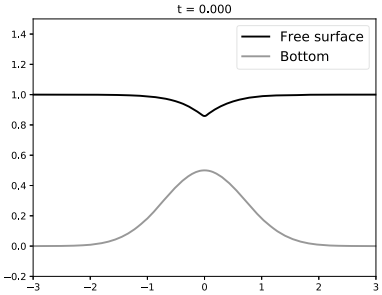
As it has been mentioned in Sect. 2 the method introduced here only preserves stationary solutions whose regime doesn't change, i.e. subcritical or supercritical stationary solutions. To do this, when a critical state is detected in the stencil $\mathcal{S}_i$, i.e. if there exists x in the stencil such that $D_f(U(x))$ is singular, the standard CWENO reconstructions is applied. The detection of critical states is performed by using a threshold ϵ: if the Froude number Fr is close to one, in the sense that $|Fr - 1| < \epsilon$, the standard CWENO reconstruction operator is applied. Otherwise, the well-balanced reconstruction operator is computed.

To check if the numerical methods behave correctly in the presence of transcritical regimes, the following test has been considered: the shallow water equations are solved in the space interval $[-3, 3]$ and the time interval $t \in [0, 20]$ with the depth function:

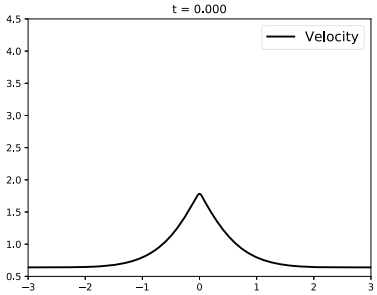
$$H(x) = -\frac{1}{2}e^{-x^2}; \tag{26}$$

As initial condition, the subcritical stationary solution of (24) satisfying $h(-3) = 1$, $q(-3) = 0.64$, is imposed (see Fig. 5).

As boundary conditions, $q(-3, t) = 1$ is set upstream, while the water height is imposed to be $h(3, t) = 1$ downstream. The CFL parameter is set to 0.5, $\Delta x = 0.02$, and $N_p = 1$ is considered. Figure 6 shows the evolution of the solution obtained with NWBM3: after the passage of the wave generated by the boundary condition, a critical state is reached at the point of minimal depth linking a subcritical regime to the left and a supercritical regime to the right, followed by a stationary hydraulic jump that links the supercritical region to the subcritical one on the right. As it can be seen the transcritical regime is well captured by the numerical method and a stationary solution is reached.

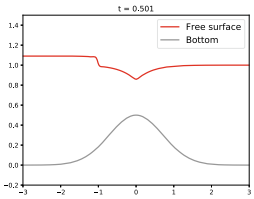


(a) Initial condition. Free surface and bottom.

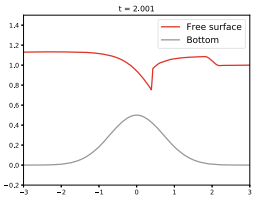


(b) Initial condition. Velocity.

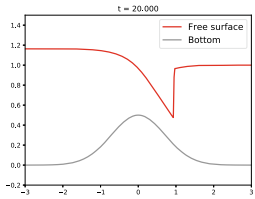
Fig. 5 Test 5.2.2. Initial condition: a subcritical stationary solution computed with RK4



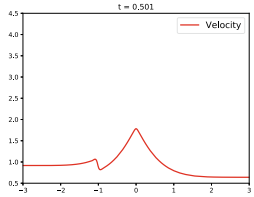
(a) Free surface. $t = 0.5s$



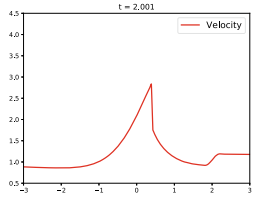
(b) Free surface. $t = 2s$



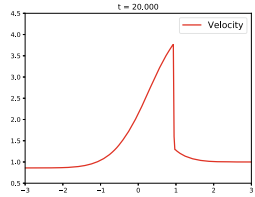
(c) Free surface. $t = 20s$



(d) Velocity. $t = 0.5s$



(e) Velocity. $t = 2s$



(f) Velocity. $t = 20s$

Fig. 6 Test 5.2.2. Evolution from a subsonic to a transonic regime simulated with NWBM3. Number of cells: 300

Therefore, this basic strategy described in this section allows us to simulate the evolution of transcritical stationary solutions.

6 Conclusions

The methodology presented in [1] has been followed to obtain a general family of high-order well-balanced numerical methods that can be applied to 1d systems of balance laws. The main difficulty related to the implementation of these methods

is that a nonlinear problem has to be solved at every cell and at every time step consisting in finding a stationary solution whose average is the given cell value. This problem has been interpreted as a control one related to an ODE system, in which the constraint is the given average and the control is the initial condition. The problem has been then written in functional form and the gradient of the functional has been computed with the help of the adjoint system. Once the expression of the gradient has been obtained, Newton's and descent methods have been applied. In order to test the efficiency of both methods, they have been tested in several problems, showing that the Newton's method is more efficient: a nonlinear scalar test problem has been shown as an example.

In order to test the efficiency and the well-balancedness of the methods, they have been applied to two problems: the Burgers equation with a nonlinear source term and the shallow water model. The tests put on evidence that the well-balanced modification increases the computational cost. In any case, this extra computational cost is lower than the one that would require to lead the discretization errors to (close to) zero machine by refining the mesh or increasing the order of non-well-balanced methods.

Further developments include applications of the introduced technique to:

- Systems of balance laws (1) in which the function H has jump discontinuities.
- Transcritical stationary solutions.
- Multidimensional problems.

References

1. Castro, M.J., Gallardo, J.M., López-García, J.A., Parés, C.: Well-balanced high order extensions of Godunov method for linear balance laws. SIAM J. Numer. Anal. **46**, 1012–1039 (2008)
2. Castro, M.J., Morales de Luna, T., Parés, C.: Well-balanced schemes and path-conservative numerical methods. In: Abgrall, R., Shu, C.-W. (eds.) Handbook of Numerical Methods for Hyperbolic Problems, vol. 18, pp. 131–175. Elsevier, Amsterdam (2017)
3. Cravero, I., Semplice, M.: On the accuracy of WENO and CWENO reconstructions of third order on nonuniform meshes. J. Sci. Comput. **67**, 1219–1246 (2016)
4. Castro, M.J., López-García, J.A., Parés, C.: In high order exactly well-balanced numerical methods for shallow water systems. J. Comput. Phys. **246**, 242–264 (2013)
5. Castro, M.J., Parés, C.: Well-balanced high-order finite volume methods for systems of balance laws. J. Sci. Comput. **82**, 48 (2020)
6. Gómez-Bueno, I., Castro, M.J., Parés, C.: High-order well-balanced methods for systems of balance laws: a control-based approach **394**, 125820 (2021)
7. Gottlieb, S., Shu, C.-W.: Total variation diminishing Runge-Kutta schemes. Math. Comput. **67**, 73–85 (1998)
8. Hager, W.W., Xhang, H.: A survey of nonlinear conjugate gradient methods. Pac. J. Optimiz. **2**, 35–58 (2006)
9. Levy, D., Puppo, G., Russo, G.: Compact central WENO schemes for multidimensional conservation laws. SIAM J. Sci. Comput. **22**, 656–672 (2000)

On High-Precision L^∞-stable IMEX Schemes for Scalar Hyperbolic Multi-scale Equations

Victor Michel-Dansac and Andrea Thomann

Abstract We present a framework to build high-accuracy IMEX schemes that fulfill the maximum principle, applied to a scalar hyperbolic multi-scale equation. Motivated by the findings in [5] that implicit R-K schemes are not L^∞-stable, our scheme, for which we can prove the L^∞ stability, is based on a convex combination between a first-order and a class of second-order IMEX schemes. We numerically demonstrate the advantages of our scheme, especially for discontinuous problems, and give a MOOD procedure to increase the precision.

Keywords L^∞ stability · IMEX R-K schemes · MOOD · Hyperbolic multi-scale equations

1 Introduction

We consider the scalar multi-scale equation

V. Michel-Dansac (✉)
Institut de Mathématiques de Toulouse, Université Toulouse 3 Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 9, France

INSA Toulouse, 135 Avenue de Rangueil, 31077 Toulouse Cedex 4, France

Université de Strasbourg, CNRS, Inria, IRMA, 67000 Strasbourg, France
e-mail: victor.michel-dansac@inria.fr

A. Thomann
Dipartimento di Scienze e Alta Tecnologia, Università degli Studi dell'Insubria, Via Valleggio 11, 22100 Como, Italy

Marie Skłodowska-Curie fellow of the Istituto Nazionale di Alta Matematica Francesco Severi, Rome, Italy

Institute of Mathematics, Johannes Gutenberg University Mainz, Staudingerweg 9, 55128 Mainz, Germany
e-mail: athomann@uni-mainz.de

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*, SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2_4

$$w_t + c_e w_x + \frac{c_i}{\varepsilon} w_x = 0, \tag{1}$$

where we set the constants $c_e, c_i > 0$ and the parameter $\varepsilon > 0$. The model (1) mimics the behavior of the isentropic Euler equations with a slow speed c_e and a fast speed c_i/ε, where ε corresponds to the square of the Mach number. We treat the derivative w_x associated with the slow scale c_e explicitly, whereas w_x associated with the fast scale c_i/ε is treated implicitly in time due to the stiffness introduced by $\varepsilon < 1$. For computational efficiency, the resulting CFL condition, and therefore the time step, has to be independent of ε. In space, we apply an upwind discretization because, already having in mind the non-linear nature of e.g. the Euler equations, using a central scheme for the implicit part will not lead to a L^∞-stable scheme, as shown in [4] for a non-linear system.

The discretization of time and space follows the usual finite difference framework. The space domain $[x_1, x_N]$ is partitioned in N uniformly spaced points $(x_j)_{j\in[1,N]}$, with the step size Δx. We discretize the time variable with $t^n = n\Delta t$, where Δt denotes the time step. Then, the solution $w(t, x)$ of (1) at (t^n, x_j) is approximated by w_j^n. The first-order implicit-explicit (IMEX) discretization of (1) is given by

$$w_j^{n+1} = w_j - \lambda(w_j^n - w_{j-1}^n) - \mu_\varepsilon(w_j^{n+1} - w_{j-1}^{n+1}), \tag{2}$$

where we define $\lambda = c_e \frac{\Delta t}{\Delta x}$ and $\mu_\varepsilon = \frac{c_i}{\varepsilon}\frac{\Delta t}{\Delta x}$ for abbreviation. Note that $\lambda, \mu_\varepsilon > 0$.

We are interested in IMEX schemes that meet the maximum principle. Here, we focus on L^∞-stable schemes, where a scheme is said to be L^∞-stable if

$$\|w^{n+1}\|_\infty = \max_{j\in[\![1,N]\!]} |w_j^{n+1}| \leq \|w^n\|_\infty. \tag{3}$$

As proven in [3], the first-order scheme (2) is L^∞-stable and TVD. Furthermore, as proven in [5], implicit Runge-Kutta schemes, and consequently second-order IMEX schemes, are not L^∞-stable. Therefore, we would like to propose a convex combination of (2) with a second-order IMEX scheme and give conditions for the L^∞ stability for the resulting scheme. We define the convex combination between the first-order scheme $w_j^{n+1,1\text{st}}$ and a second-order update $w_j^{n+1,2\text{nd}}$ for a parameter $\theta \in (0, 1)$ as:

$$w_j^{n+1} = (1-\theta)\, w_j^{n+1,1\text{st}} + \theta\, w_j^{n+1,2\text{nd}}. \tag{4}$$

2 IMEX Formulation

Generic formulations of an IMEX scheme introduce two $s \times s$ matrices $A = (a_{ij})$ and $\tilde{A} = (\tilde{a}_{ij})$, as well as two vectors $b, \tilde{b} \in \mathbb{R}^s$. They are regrouped in two linked Butcher tableaux

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}, \qquad \begin{array}{c|c} \tilde{c} & \tilde{A} \\ \hline & \tilde{b}^T \end{array}.$$

The coefficients $c, \tilde{c}$ are only necessary if the right hand side depends explicitly on time. In the following we will use the pairs (A, b) for the implicit and $(\tilde{A}, \tilde{b})$ for the explicit part. To reduce computational costs, we take A to be lower triangular and $\tilde{A}$ to be strictly lower triangular. Applying the IMEX formulation on (1), we obtain the following scheme:

$$w^{n+1} = w^n - c_e \Delta t \sum_{k=1}^{s} \tilde{b}_k w_x^{(k)} - \frac{c_i}{\varepsilon} \Delta t \sum_{k=1}^{s} b_k w_x^{(k)}, \tag{5}$$

with the stages

$$w^{(k)} = w^n - c_e \Delta t \sum_{l=1}^{k-1} \tilde{a}_{kl} w_x^{(l)} - \frac{c_i}{\varepsilon} \Delta t \sum_{l=1}^{k} a_{kl} w_x^{(l)}. \tag{6}$$

IMEX Runge-Kutta (R-K) schemes can be classified depending on the structure of the matrix A.

Definition 1 An IMEX R-K method is said to be of type CK (Carpenter and Kennedy [6]) if the matrix $A \in \mathbb{R}^{s\times s}$ can be written as

$$A = \begin{pmatrix} 0 & 0 \\ a & \hat{A} \end{pmatrix},$$

where $a \in \mathbb{R}^{s-1}$ and $\hat{A} \in \mathbb{R}^{(s-1)\times(s-1)}$ is invertible. In the case where $a = 0$, the scheme is said to be of ARS type (Asher, Ruuth and Spiteri [1]).

In the following we will consider a second-order 2-stage and a second-order 3-stage IMEX R-K method of type CK. To obtain a second-order scheme, there are the following compatibility conditions [9]:

$$\begin{aligned} &\sum_{k=1}^{s} \tilde{b}_k = 1; \quad \sum_{k=1}^{s} b_k = 1; \quad \forall k,\ \tilde{c}_k = \sum_{l=1}^{k-1} \tilde{a}_{kl}; \quad \forall k,\ c_k = \sum_{l=1}^{k-1} a_{kl}\,; \\ &\sum_{k=1}^{s} \tilde{b}_k \tilde{c}_k = \frac{1}{2}; \quad \sum_{k=1}^{s} \tilde{b}_k c_k = \frac{1}{2}; \quad \sum_{k=1}^{s} b_k \tilde{c}_k = \frac{1}{2}; \quad \sum_{k=1}^{s} b_k c_k = \frac{1}{2}. \end{aligned} \tag{7}$$

2.1 A 2-Stage CK Type IMEX R-K Method

For a 2-stage CK type method, we have the following Butcher tableaux, with $a_{22} \neq 0$:

$$\text{explicit: } \begin{array}{c|cc} 0 & 0 & 0 \\ \tilde{c}_2 & \tilde{a}_{21} & 0 \\ \hline & \tilde{b}_1 & \tilde{b}_2 \end{array}, \qquad \text{implicit: } \begin{array}{c|cc} 0 & 0 & 0 \\ c_2 & a_{21} & a_{22} \\ \hline & b_1 & b_2 \end{array}. \tag{8}$$

With the compatibility conditions (7), we can simplify (8) to

$$\text{explicit: } \begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}, \qquad \text{implicit: } \begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \gamma & \alpha-\gamma \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}, \tag{9}$$

where $\alpha - \gamma \neq 0$ and $\alpha \neq 0$.

Using (5), (6) and (9), we can define the second-order discretization of (1) as

$$\begin{aligned} w_j^{(1)} &= w_j^n - \lambda\alpha(w_j^n - w_{j-1}^n) - \gamma\mu_\varepsilon(w_j^n - w_{j-1}^n) - \mu_\varepsilon(\alpha-\gamma)(w_j^{(1)} - w_{j-1}^{(1)}), \\ w_j^{n+1} &= w_j^n - \left(1 - \frac{1}{2\alpha}\right)(\lambda + \mu_\varepsilon)(w_j^n - w_{j-1}^n) - \frac{1}{2\alpha}(\lambda+\mu_\varepsilon)(w_j^{(1)} - w_{j-1}^{(1)}). \end{aligned} \tag{10}$$

Due to the matrix structure of the CK type R-K scheme, we have only two computational steps. The first one computes $w^{(1)}$, and the second one w^{n+1}. The convex combination (4) between the schemes (2) and (10), with the shorter notation $\Delta = w_j - w_{j-1}$, is given by:

$$\begin{aligned} w_j^{(1)} &= w_j^n - \lambda\alpha\Delta^n - \gamma\mu_\varepsilon\Delta^n - \mu_\varepsilon(\alpha-\gamma)\Delta^{(1)}, \\ w_j^{n+1} &= w_j^n - \left(\lambda - \theta\frac{1}{2\alpha}(\lambda+\mu_\varepsilon) + \theta\mu_\varepsilon\right)\Delta^n \\ &\quad - \theta\frac{1}{2\alpha}(\lambda+\mu_\varepsilon)\Delta^{(1)} - (1-\theta)\mu_\varepsilon\Delta^{n+1}. \end{aligned} \tag{11}$$

We can sort (11) by grouping the w^{n+1} and $w^{(1)}$ terms:

$$\begin{aligned} (1+\mu_\varepsilon(\alpha-\gamma))w_j^{(1)} - \mu_\varepsilon(\alpha-\gamma)w_{j-1}^{(1)} &= (1-(\lambda\alpha+\gamma\mu_\varepsilon))w_j^n \\ &\quad + (\lambda\alpha+\gamma\mu_\varepsilon)w_{j-1}^n, \end{aligned} \tag{12}$$

$$\begin{aligned} (1+(1-\theta)\mu_\varepsilon)w_j^{n+1} - (1-\theta)\mu_\varepsilon w_{j-1}^{n+1} &= w_j^n - \left(\lambda - \theta\frac{1}{2\alpha}(\lambda+\mu_\varepsilon) + \theta\mu_\varepsilon\right)\Delta^n \\ &\quad - \theta\frac{1}{2\alpha}(\lambda+\mu_\varepsilon)\Delta^{(1)}. \end{aligned} \tag{13}$$

In the following, we will assume periodic boundary conditions. We will prove the L^∞ stability (3) by starting with the proof of $\|w^{(1)}\|_\infty \leq \|w^n\|_\infty$. For each time step, we

will use the triangle inequality $|x+y| \le |x| + |y|$ and the reverse triangle inequality $|x| - |y| \le |x-y|$ for $x, y \in \mathbb{R}$. To apply use those inequalities, we require in (12)

$$\lambda\alpha + \gamma\mu_\varepsilon \ge 0 \tag{14}$$

$$1 - (\lambda\alpha + \gamma\mu_\varepsilon) \ge 0 \tag{15}$$

$$1 + \mu_\varepsilon(\alpha - \gamma) \ge 0 \tag{16}$$

$$\mu_\varepsilon(\alpha - \gamma) \ge 0. \tag{17}$$

Using equation (12), we can now write the following estimate for $\|w^n\|_\infty$:

$$\begin{aligned}
\|w^n\|_\infty &= (1 - (\lambda\alpha + \gamma\mu_\varepsilon))\|w^n\|_\infty + (\lambda\alpha + \gamma\mu_\varepsilon)\|w^n\|_\infty \\
&\ge \|(1 - (\lambda\alpha + \gamma\mu_\varepsilon))w_j^n + (\lambda\alpha + \gamma\mu_\varepsilon)w_{j-1}^n\|_\infty \\
&= \|(1 + \mu_\varepsilon(\alpha - \gamma))w_j^{(1)} - \mu_\varepsilon(\alpha - \gamma)w_{j-1}^{(1)}\|_\infty \\
&\ge (1 + \mu_\varepsilon(\alpha - \gamma))\|w^{(1)}\|_\infty - \mu_\varepsilon(\alpha - \gamma)\|w^{(1)}\|_\infty \\
&= \|w^{(1)}\|_\infty.
\end{aligned}$$

From requirement (14), we get that $\alpha c_e + \gamma \frac{c_i}{\varepsilon} \ge 0$. In order to get a Butcher tableau independent of ε, we require $\alpha > 0$ and $\gamma \ge 0$. Relation (15) leads to a CFL condition $\frac{\Delta t}{\Delta x}(\alpha c_e + \gamma \frac{c_i}{\varepsilon}) \le 1$. Note that, due to computational efficiency, we seek a time step restriction independent of ε. Therefore, we must take $\gamma = 0$, which is compatible with the restriction $\gamma \ge 0$. With those settings, (16) and (17) are always fulfilled.

Let us prove now that $\|w^{n+1}\|_\infty \le \|w^n\|_\infty$. First, we rewrite (12) as follows:

$$-\mu_\varepsilon \Delta^{(1)} = \frac{1}{\alpha} w_j^{(1)} - \frac{1}{\alpha} w_j^n + \lambda(w_j^n - w_{j-1}^n). \tag{18}$$

After inserting (18) into (13), we obtain further conditions given by

$$r_1 = 1 - \frac{\theta}{2\alpha^2} + \lambda\left(-1 + \frac{\theta}{\alpha}\right) + \mu_\varepsilon\theta\left(-1 + \frac{1}{2\alpha}\right) \ge 0, \tag{19}$$

$$r_2 = \lambda\left(1 - \frac{\theta}{\alpha}\right) + \mu_\varepsilon\theta\left(1 - \frac{1}{2\alpha}\right) \ge 0, \tag{20}$$

$$\frac{\theta}{2\alpha^2} - \frac{\theta\lambda}{2\alpha} \ge 0. \tag{21}$$

Using (13), as well as the above conditions, we obtain the following estimate

$$\begin{aligned}
\|w^{n+1}\|_\infty &= (1+(1-\theta)\mu_\varepsilon)\|w^{n+1}\|_\infty - (1-\theta)\mu_\varepsilon\|w^{n+1}\|_\infty \\
&\leq \|(1+(1-\theta)\mu_\varepsilon)w_j^{n+1} - (1-\theta)\mu_\varepsilon w_{j-1}^{n+1}\|_\infty \\
&= \|r_1 w_j^n + r_2 w_{j-1}^n + (\mu_\varepsilon - \frac{\theta\lambda}{2a})w_j^{(1)} + \frac{\theta\lambda}{2\alpha}w_{j-1}^{(1)}\|_\infty \\
&\leq \left(1 - \frac{\theta}{2\alpha^2}\right)\|w^n\|_\infty + \frac{\theta}{2\alpha^2}\|w^{(1)}\|_\infty \\
&\leq \|w^n\|_\infty,
\end{aligned}$$

which shows the L^∞ stability. From the constraints (19)–(21), we can compute the final estimates for the free parameters α, θ, λ. The condition (21) gives a CFL restriction of $\lambda \leq \frac{1}{\alpha}$. Since we want to avoid a dependence of c_e, c_i or ε on α and θ, we need in (20) $1 - \frac{\theta}{\alpha} \geq 0$, that is $\alpha \geq \theta$ and $1 - \frac{1}{2a} \geq 0$, which leads to $\alpha \geq \frac{1}{2}$. With the same motivation, we need $-1 + \frac{1}{2\alpha} \geq 0$ in (19), that is $\alpha \leq \frac{1}{2}$. Together it follows that $\alpha = \frac{1}{2}$ and we get from (19) the final CFL condition $\lambda \leq 1$. With $\alpha = \frac{1}{2}$ and $\gamma = 0$ fixed, we have recovered a 2-stage ARS type method with the midpoint rule as the implicit part, given by

$$\text{explicit: } \begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}, \qquad \text{implicit: } \begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \hline & 0 & 1 \end{array}. \tag{22}$$

The above results are summed up in the following theorem:

Theorem 1 *For periodic boundary conditions and under the CFL condition*

$$\Delta t \leq \frac{\Delta x}{c_e},$$

the scheme consisting of the convex combination of the first-order scheme (2) *and the second-order scheme constructed from* (22) *is* L^∞*-stable as long as* $\theta \leq \frac{1}{2}$.

Remark 1 In order to have the maximal input of the second-order scheme, we would want to set $\theta = \theta_{\text{opt}} = \frac{1}{2}$. With this choice of θ, the restriction (19) for $\alpha = \frac{1}{2}$ is satisfied immediately and we get the less restrictive CFL condition

$$\Delta t \leq 2\frac{\Delta x}{c_e}.$$

Unfortunately, the midpoint rule with the above CFL condition and $\theta = \theta_{\text{opt}}$ exactly reduces to two steps of a first-order scheme. We therefore advise $\theta < \frac{1}{2}$ to get a second-order scheme.

Since $\gamma = 0$, the initial CK type method (9) reduces to an ARS type method (22). This observation is summarized in the following corollary

Corollary 1 *If there is a second-order CK type IMEX R-K scheme of the form (9) that is L^∞-stable in the convex combination with (2) under a CFL condition independent of ε, then it has to be of ARS type.*

2.2 A 3-Stage CK Type IMEX R-K Method

In this section, we adapt the derivation of the 2-stage case to a 3-stage CK type method. It is described by the following Butcher tableaux, with $a_{22} \neq 0$ and $a_{33} \neq 0$:

$$\text{explicit:}\ \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \tilde{c}_2 & \tilde{a}_{21} & 0 & 0 \\ \tilde{c}_3 & \tilde{a}_{31} & \tilde{a}_{32} & 0 \\ \hline & \tilde{a}_{31} & \tilde{a}_{32} & 0 \end{array}, \qquad \text{implicit:}\ \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ c_2 & a_{21} & a_{22} & 0 \\ c_3 & a_{31} & a_{32} & a_{33} \\ \hline & a_{31} & a_{32} & a_{33} \end{array}, \tag{23}$$

To have the same number of computational steps as in the 2-stage scheme (5), we have set $b = (a_{3j})$ and $\tilde{b} = (\tilde{a}_{3j})$.

With the second-order compatibility conditions (7) and $a_{22} = \beta$ and $a_{33} = \alpha$, we introduce $\kappa = \frac{2(\gamma+\beta)(1-\alpha)+2\alpha-1}{2(\gamma+\beta)}$ and simplify (23) to:

$$\text{explicit:}\ \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma+\beta & \gamma+\beta & 0 & 0 \\ 1 & 1-\frac{1}{2(\gamma+\beta)} & \frac{1}{2(\gamma+\beta)} & 0 \\ \hline & 1-\frac{1}{2(\gamma+\beta)} & \frac{1}{2(\gamma+\beta)} & 0 \end{array}, \qquad \text{implicit:}\ \begin{array}{c|ccc} 0 & 0 & 0 & \\ \gamma+\beta & \gamma & \beta & 0 \\ 1 & \kappa & \frac{1-2\alpha}{2(\gamma+\beta)} & \alpha \\ \hline & \kappa & \frac{1-2\alpha}{2(\gamma+\beta)} & \alpha \end{array}. \tag{24}$$

Analogously to (10), we can write the second-order scheme using (24) as

$$\begin{aligned} w_j^{(1)} + \mu_\varepsilon \beta \Delta^{(1)} &= w_j^n - (\lambda(\gamma+\beta) + \mu_\varepsilon \gamma)\Delta^n \\ w_j^{n+1} + \mu_\varepsilon \alpha \Delta^{n+1} &= w_j^n - \left(\lambda \frac{2(\gamma+\beta)-1}{2(\gamma+\beta)} + \kappa \mu_\varepsilon\right)\Delta^n \\ &\quad - \left(\lambda \frac{1}{2(\gamma+\beta)} + \mu_\varepsilon \frac{1-2\alpha}{2(\gamma+\beta)}\right)\Delta^{(1)}. \end{aligned}$$

We conduct an analogous analysis as in the 2-stage case, which results in the following ARS-type IMEX R-K method for $\beta \in (0, \frac{1}{2})$:

$$\text{explicit:}\ \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \beta & \beta & 0 & 0 \\ 1 & 1-\frac{1}{2\beta} & \frac{1}{2\beta} & 0 \\ \hline & 1-\frac{1}{2\beta} & \frac{1}{2\beta} & 0 \end{array}, \qquad \text{implicit:}\ \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \beta & 0 & \beta & 0 \\ 1 & 0 & \frac{1}{2(1-\beta)} & 1-\frac{1}{2(1-\beta)} \\ \hline & 0 & \frac{1}{2(1-\beta)} & 1-\frac{1}{2(1-\beta)} \end{array}. \tag{25}$$

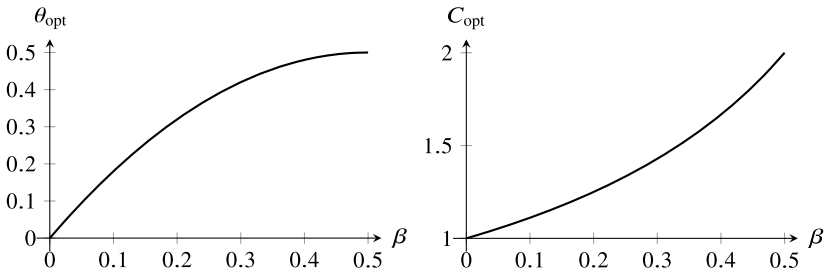


Fig. 1 Values of the optimal convex combination parameter θ_{opt} (left panel) and the optimal CFL number C_{opt} (right panel), with respect to the IMEX parameter β.

One example for (25) is the widely used ARS(2,2,2) method with $\beta = 1 - \frac{\sqrt{2}}{2}$, see [1].

Theorem 2 *For periodic boundary conditions and under the CFL condition*

$$\Delta t \leq \frac{\Delta x}{c_e},$$

the scheme consisting of the convex combination of the first-order scheme (2) *and the second-order scheme constructed from* (25) *with* $\beta \in (0, \frac{1}{2})$ *is* L^∞*-stable as long as* $\theta \leq 2\beta(1-\beta)$.

Remark 2 In order to have the maximal input of the second-order scheme, we set

$$\theta_{\text{opt}} = 2\beta(1-\beta). \tag{26}$$

With the choice $\theta = \theta_{\text{opt}}$, we get the less restrictive CFL condition

$$\Delta t \leq C_{\text{opt}} \frac{\Delta x}{c_e}, \text{ where } C_{\text{opt}} = \frac{1}{1-\beta}. \tag{27}$$

The values of θ_{opt} and C_{opt} are displayed with respect to β in Fig. 1.

Remark 3 Allowing $\beta = \frac{1}{2}$, the 3-stage ARS type method (25) reduces to the 2-stage ARS type method using the midpoint rule (22). In addition, the choice $\beta = \frac{1}{2}$ maximizes both θ_{opt} and λ.

Corollary 2 *If there is a second-order CK type IMEX R-K scheme of the form* (25) *that is* L^∞*-stable in the convex combination with* (2) *under a CFL condition independent of* ε*, then it has to be of ARS type.*

3 Numerical Results

This section is dedicated to providing numerical experiments to test the schemes introduced above:

- The first-order scheme given by (2),
- The second-order scheme given by (25),
- The L^∞-stable scheme obtained via the convex combination with the parameter $\theta = \theta_{\text{opt}}$ given by (26), between the first-order scheme (2) and the second-order scheme (25),
- The MOOD scheme resulting from an optimal order detection procedure explained in Sect. 3.1 and applied to the L^∞-stable scheme.

Throughout this section, the space domain is given by $[0, 1]$ and periodic boundary conditions are prescribed. The time domain is given by $[0, t_{\text{end}}]$, where t_{end} chosen such that the exact solution completes exactly one revolution of the space domain, as follows:

$$t_{\text{end}} = \frac{1}{c_e + \frac{c_i}{\varepsilon}}.$$

Unless otherwise mentioned, the space and time discretizations are linked with the optimal CFL condition defined by (27). The constants c_e and c_i are both taken equal to 1.

We start this section with an introduction to an order detection procedure in Sect. 3.1. Then, we provide a way to choose the parameter β in Sect. 3.2. Finally, in Sect. 3.3, we provide several numerical tests with smooth and especially non-smooth exact solutions. The smooth exact solution is given by

$$w_{\text{ex}}^{\text{smooth}}(t, x) = 1 + \frac{\varepsilon}{2}\left(1 + \sin\left(2\pi\left(x - \left(c_e + \frac{c_i}{\varepsilon}\right)t\right)\right)\right), \tag{28}$$

and describes the transport of a sine wave of amplitude ε. The discontinuous exact solution is given by

$$w_{\text{ex}} = \begin{cases} 1 + \varepsilon & \text{if } x - \left(c_e + \frac{c_i}{\varepsilon}\right)t \in \left(\frac{1}{4}, \frac{3}{4}\right), \\ 1 & \text{otherwise.} \end{cases} \tag{29}$$

which corresponds to the transport of a square wave of amplitude ε.

3.1 Optimal Order Detection: A MOOD-like Technique

The L^∞-stable scheme is a convex combination between the diffusive first-order scheme and the oscillatory second-order scheme. Since those oscillations may violate

the maximum principle, we do not wish to use the second-order scheme everywhere in the computational domain. Using the L^∞-stable scheme introduces enough diffusion to get rid of the oscillations and to ensure the maximum principle. However, once the diffusion has been introduced, there is no need to add even more diffusion and the second-order scheme could be used until its result once again violates the maximum principle, at which point the L^∞-stable scheme is necessary once again.

The procedure outlined above is akin to the Multidimensional Optimal-Order Detection techniques developed in the MOOD framework (see for instance [2]). It results in the MOOD scheme, given by the algorithm below:

Algorithm If the exact solution is bounded between $w_{\min}$ and $w_{\max}$, using the optimal CFL number (27), the MOOD scheme is given as a result of applying the following algorithm at each time step.

1. Compute the second-order solution.
2. Detect if this second-order solution breaks the maximum principle, i.e. if it oscillates below $w_{\min}$ or above $w_{\max}$.
3. If the maximum principle is violated, compute and output the solution given by the L^∞-stable scheme; otherwise, output the second-order solution.

This algorithm ensures a drastic improvement in the numerical results when this procedure is used instead of using the L^∞-stable scheme at each time step.

3.2 Choice of β in the 3-Stage Method

This first set of numerical experiments is dedicated to providing a way to choose an optimal value for β. At the moment, we know that $\beta \in (0, 1/2)$ and we are able to find a non-zero value of θ for all values of β. According to Fig. 1, the optimal CFL number as well as the optimal θ increase as β goes to $1/2$. Therefore, it would be tempting to take β as close to $1/2$ as possible. To check whether this preliminary analysis is accurate, we study the CPU time and the L^∞ error of the scheme with respect to β, in order to suggest an optimal value of β.

Throughout this set of numerical experiments, we consider the smooth exact solution (28) with $\varepsilon = 10^{-1}$.

Study of the CPU time. The CPU time taken by our program is influenced by β because the CFL number C_{opt}, given by (27), itself depends on β. Indeed, as β varies from 0 to $1/2$, C_{opt} ranges between 1 and 2, as evidenced in Fig. 1.

In Fig. 2, we note that the CPU time for the L^∞-stable and MOOD schemes decreases when β tends to $1/2$. This was expected as the CFL number C_{opt} is increasing with β, thus allowing for larger time steps. Let us also note that the MOOD procedure is not very costly for this smooth test case. Moreover, we remark that the second-order scheme takes twice as much CPU time as the first-order scheme, which is also expected due to the additional intermediate step.

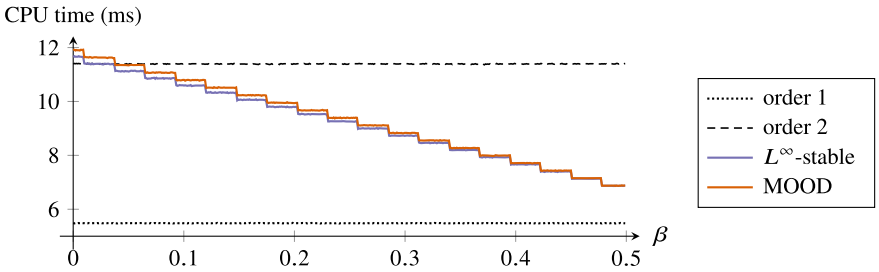


Fig. 2 CPU time (in milliseconds) with respect to the IMEX parameter β, using the optimal values θ_{opt} and C_{opt}, in the context of the test case presented in Sect. 3.2.

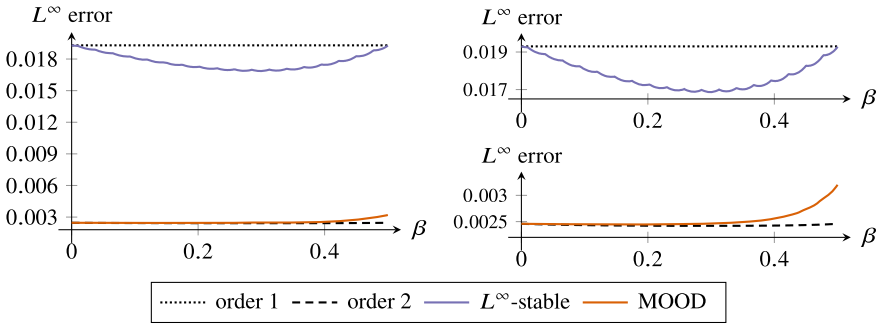


Fig. 3 L^∞ error with respect to the IMEX parameter β, using the optimal values θ_{opt} and C_{opt}, in the context of the test case presented in Sect. 3.2. The right panels contain a zoom on the left panel data.

Study of the L^∞ Error. Now, we turn to the study of the L^∞ error with respect to β. For $\beta \in (0, 1/2)$, the L^∞-stable and MOOD schemes are L^∞-stable, but this property alone does not indicate their precision. From now on, we take the optimal CFL number C_{opt}.

In Fig. 3, we observe that the second-order scheme is, as expected, much more precise than the first-order one. In addition, we note that the L^∞-stable scheme is more precise than the first-order one, but not by a large margin. Finally, we remark that the MOOD procedure is essential to improve the precision of the L^∞-stable scheme.

Regarding the choice of β, we note on the top right panel that the L^∞-stable scheme reduces to the first-order one in two cases. When $\beta = 0$, we get $\theta_{\text{opt}} = 0$, and the convex combination consists only in the first-order scheme. When $\beta = 1/2$, we get $\theta_{\text{opt}} = 0$ and $C_{\text{opt}} = 2$, and the convex combination actually coincides with the first-order scheme. Between these two values, the L^∞ error produced by the L^∞-stable scheme reaches a minimum. Interestingly, this minimum is close to the point where the MOOD error starts increasing (see the bottom right panel). We note that this minimum is located around $\beta \simeq 1 - \sqrt{2}/2$, which is widely used e.g. in [1, 3, 9].

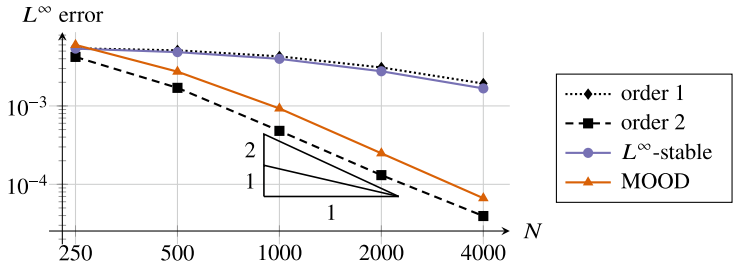


Fig. 4 L^∞ error curves for the smooth solution (28), with $\varepsilon = 10^{-2}$.

Conclusion: Choice of β. In this first study, we have observed that:

- the CPU time gets smaller as β gets larger;
- the L^∞ error reaches a minimum at $\beta = 1 - \sqrt{2}/2$.

Based on this observations, we define β_{opt}, which will be used in the remainder of this article, as

$$\beta_{\text{opt}} = 1 - \frac{\sqrt{2}}{2}.$$

3.3 Numerical Tests

Before we start with the numerical results, we want to remark that we do not consider an increase in space order. Such an increase, and its effect on smooth solutions, has been documented at length in [3]. Therein was concluded that, if ε is close to 1, then using a second-order scheme in time and a first-order scheme in space does not provide a significant and observable gain compared to a first-order scheme in time and space. Conversely, if ε is close to 0, then using a first-order scheme in time and a second-order scheme in space does not provide a significant and observable gain compared to a first-order scheme in time and space.

Therefore, we focus here only on second-order time accuracy whereas accuracy in space will be studied in forthcoming work.

3.3.1 Smooth Solution: Order of Accuracy

To demonstrate that our schemes reach the desired order of accuracy, we compute L^∞ error curves with the smooth initial condition (28). In Fig. 4, we display the L^∞ error with respect to the number of discretization points for the four schemes under consideration.

We note, as expected, that the first- and second-order schemes are respectively first- and second-order accurate. Moreover, the L^∞-stable scheme is first-order accu-

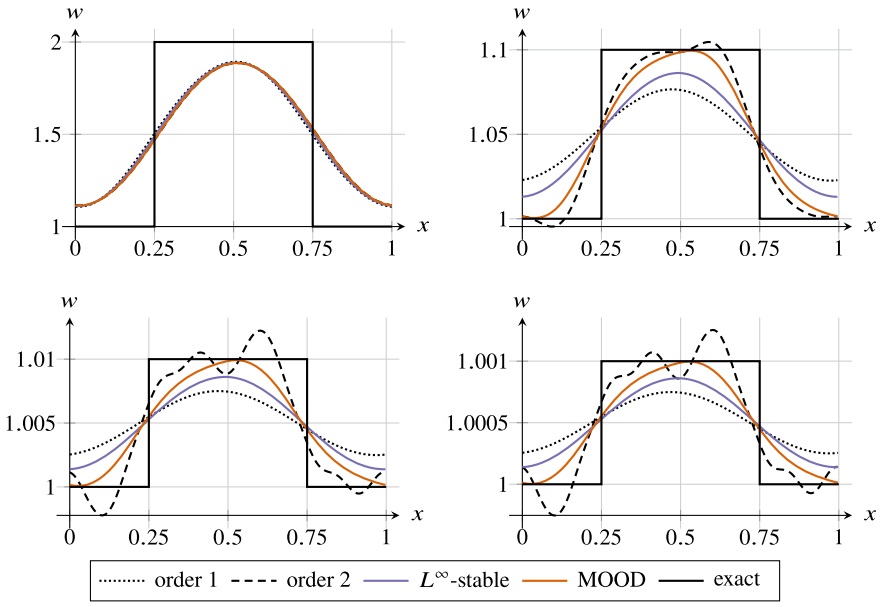


Fig. 5 Approximation of the discontinuous solution (29). From left to right and top to bottom, we have taken: $\varepsilon = 1$ and $N = 40$, $\varepsilon = 10^{-1}$ and $N = 220$, $\varepsilon = 10^{-2}$ and $N = 2000$, $\varepsilon = 10^{-3}$ and $N = 20000$. These large values of N have been chosen to ensure that 20 time iterations are systematically needed to reach t_{end}. If smaller values are taken, the time steps are too large to visualize noticeable differences between the schemes.

rate, and the MOOD procedure greatly increases the precision of the L^∞-stable scheme, almost bringing it to the level of the second-order scheme. The loss of precision of the MOOD scheme compared to the second-order scheme is due to the fact that the MOOD scheme is L^∞-stable, contrary to the second-order scheme, and therefore it does not allow any violation of the maximum principle, even if such a violation would result in a precision increase.

As a consequence, the MOOD procedure is especially well-suited for smooth problems where the maximum principle is important. Let us now compare these approaches on a discontinuous solution, where we expect the L^∞-stable scheme to be of greater interest.

3.3.2 Discontinuous Solution

We now consider the following discontinuous exact solution w_{ex}. In Fig. 5, we display the results of the four schemes for different values of ε.

We first notice in the top left panel that the approximation of the exact solution is similar for all four schemes in the case of $\varepsilon = 1$.

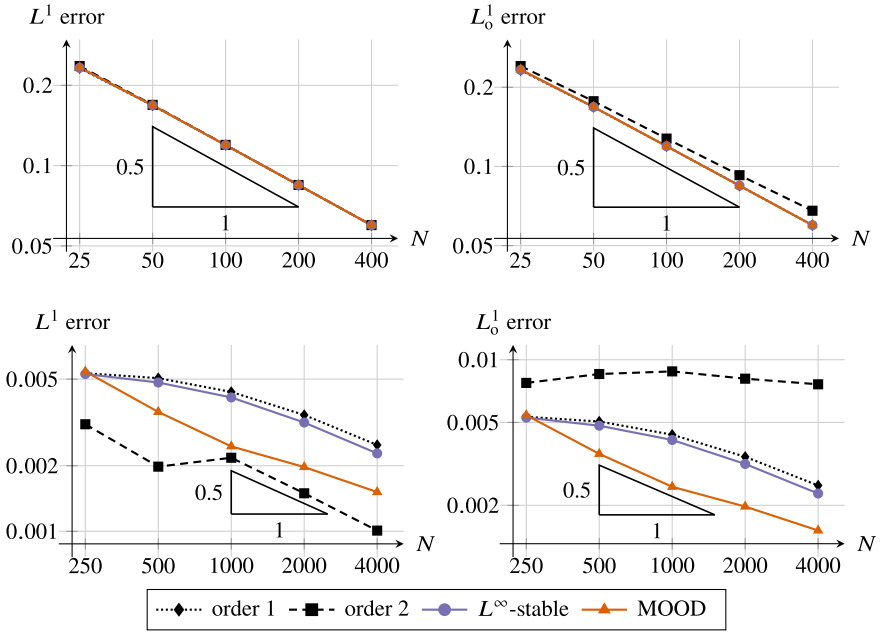


Fig. 6 L^1 (left panels) and L^1_o (right panels) error curves for the discontinuous solution (29), for $\varepsilon = 1$ (top panels) and $\varepsilon = 10^{-2}$ (bottom panels).

In the other three panels, for $\varepsilon \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, we note that the first-order scheme is always in-bounds, while the second-order scheme always violates the maximum principle. Here, we observe a clear improvement when using the L^∞-stable scheme, but the result is still somewhat diffusive. The MOOD procedure allows another gain in precision compared to the first-order scheme, while still staying in-bounds.

This underlines the necessity of L^∞-stable schemes when approximating discontinuous solutions. In addition, the MOOD procedure is useful when approximating continuous and discontinuous solutions with good precision, while respecting the maximum principle.

The final numerical experiment consists in quantifying how much better the result of the L^∞-stable scheme is, compared to both first- and second-order approximations, when considering a discontinuous solution. To address such an issue, we cannot simply compute the error in the L^∞ norm. Indeed, this norm is not well-suited for measuring the errors produced when approximating a discontinuous exact solution with a diffusive approximation. Instead, we turn to the L^1 norm, as well as a modification, the L^1_o quasinorm, which does not satisfy the triangle inequality property of a norm but enables us to measure relevant errors, defined as follows:

$$\|w^n\|_{L^1_o} = \frac{1}{\Delta x} \sum_j \left(|w_j^n| + \max_{m \leq n} \left[\left(\max_j w_j^m - \min_j w_j^m \right) - \left(\max_j w_j^0 - \min_j w_j^0 \right) \right] \right).$$

This quasinorm is the L^1 norm added to a quantity which has been designed to measure only overshoots and undershoots. This quantity encodes how much the numerical solution violates the maximum principle. Therefore, we expect this added term to vanish as soon as the L^∞-stable scheme, with or without MOOD, is employed.

In the top panels of Fig. 6, we note that, for $\varepsilon = 1$, both errors take similar values for the four schemes under consideration. This is due to the fact that there are few spurious oscillations in this case (see Fig. 5, top left panel). In addition, we observe that the scheme is accurate up to order $1/2$ which is expected when approximating discontinuous solutions, see for instance [7].

Now, looking at the bottom left panel, we note that the L^1 error is lower for the second-order scheme than for the other ones and that the orders of accuracy of all schemes tend to $1/2$ for large enough N. However, the bottom right panel, which takes into account the over- and undershoots when computing the error, paints another picture: the second-order scheme is actually the worst of all four. In addition, the error actually stays roughly constant when the number of discretization points increases. This means that, as N increases, the gains in L^1 error seem to be compensated by an increase of the overshoot and undershoot amplitude.

4 Conclusions and Future Work

We have presented a way of constructing L^∞-stable IMEX schemes that, combined with a MOOD procedure, yield high-precision approximate solutions for stiff and non-stiff systems. As we have demonstrated with simple numerical examples, for non-stiff systems higher order IMEX R-K schemes still give good results although violating the maximum principle, whereas for stiff systems they produce spurious oscillations and L^∞-stable schemes are needed to give accurate solutions. In this work, we have mainly focused on the time accuracy and have neglected higher order space discretizations. This, together with the extension to TVD and higher order IMEX schemes, is explored in [8]. In addition, for physical applications, asymptotic preservation properties, as well as scale-independent diffusion, will be studied.

Acknowledgments V. Michel-Dansac extends his thanks to the Service d'Hydrographie et d'Océanographie de la Marine (SHOM) for financial support. A. Thomann acknowledges the support of the INDAM-DP-COFUND-2015, grant number 713485.

References

1. Ascher, U.M., Ruuth, S.J., Spiteri, R.J.: Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. Appl. Numer. Math. **25**(2-3), 151–167 (1997). Special issue on time integration (Amsterdam, 1996)

2. Clain, S., Diot, S., Loubère, R.: A high-order finite volume method for systems of conservation laws–Multi-dimensional Optimal Order Detection (MOOD). J. Comput. Phys. **230**(10), 4028–4050 (2011)
3. Dimarco, G., Loubère, R., Michel-Dansac, V., Vignal, M.-H.: Second-order implicit-explicit total variation diminishing schemes for the Euler system in the low Mach regime. J. Comput. Phys. **372**, 178–201 (2018)
4. Dimarco, G., Loubère, R., Vignal, M.-H.: Study of a new asymptotic preserving scheme for the Euler system in the low mach number limit. SIAM J. Sci. Comput. **39**(5), A2099–A2128 (2017)
5. Gottlieb, S., Shu, C.-W., Tadmor, E.: Strong stability-preserving high-order time discretization methods. SIAM Rev. **43**(1), 89–112 (2001)
6. Kennedy, C.A., Carpenter, M.H.: Additive Runge-Kutta schemes for convection–diffusion–reaction equations. Appl. Numer. Math. **44**(1–2), 139–181 (2003)
7. LeVeque, R.J.: Numerical Methods for Conservation Laws, 2nd edn. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel (1992)
8. Michel-Dansac, V., Thomann, A.: TVD IMEX Runge-Kutta schemes based on arbitrarily high order Butcher tableaux (2020, submitted)
9. Pareschi, L., Russo, G.: Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations. In: Recent Trends in Numerical Analysis, vo. 3. Advance Theory Computational Mathematics, pp. 269–288. Nova Sci. Publ., Huntington (2001)

Numerical Methods for Specific Problems

A Staggered Pressure Correction Numerical Scheme to Compute a Travelling Reactive Interface in a Partially Premixed Mixture

D. Grapsas, R. Herbin, J.-C. Latché, and Y. Nasseri

Abstract We address a turbulent deflagration model with a flow governed by the compositional Euler equations and the flame propagation represented by the transport of the characteristic function. The numerical scheme works on staggered, unstructured meshes with a time-marching algorithm solving first the chemical species mass balances and then the mass, momentum and energy balances. A pressure correction technique is used for this latter step, which solves a balance equation for the sensible enthalpy with corrective terms to ensure consistency of the total energy. The approximate solutions respect the physical bounds and satisfy a conservative weakly-consistent discrete total energy balance equation. Numerical evidence shows that they converge to the solution of the infinitely fast chemistry continuous problem when the chemical time scale tends to zero with the space and time steps.

1 Problem Position

In this paper, we study a numerical scheme for the computation of large scale turbulent deflagrations occurring in a partially premixed atmosphere. In usual situations, such a physical phenomena is driven by the progress in the atmosphere of a shell-shaped

D. Grapsas · R. Herbin (✉) · Y. Nasseri
Aix-Marseille Université, CNRS, Centrale Marseille, I2M, UMR 7373, 13453 Marseille, France
e-mail: raphaele.herbin@univ-amu.fr

D. Grapsas
e-mail: dionysis.grapsas@ansys.com

Y. Nasseri
e-mail: youssouf.nasseri@univ-amu.fr

J.-C. Latché
Institut de Radioprotection et de Sûreté Nucléaire (IRSN), BP 3,
13115 Saint-Paul-lez-Durance cedex, France
e-mail: jean-claude.latche@irsn.fr

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*,
SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2_5

thin zone, where the chemical reaction occurs and which thus separates the burnt area from fresh gases; this zone is called the flame brush. The onset of the chemical reaction is due to the temperature elevation, so the displacement of the flame brush is driven by the heat transfer inside and in the vicinity of this zone. Modelling of deflagrations still remains a challenge, since the flame brush has a very complex structure (sometimes presented as fractal in the literature), due to thermo-convective instabilities or turbulence [14, 16]. Whatever the modelling strategy, the problem thus needs a multiscale approach, since the local flame brush structure is out of reach of the computations aimed at simulating the flow dynamics at the observation scale, *i.e.* the whole reactive atmosphere scale. A possible way to completely circumvent this problem is to perform an explicit computation of the flame brush location, solving a transport-like equation for a characteristic function of the burnt zone; such an approach transfers the modelling difficulty to the evaluation of the flame brush velocity (or, more precisely speaking, to the relative velocity of the flame brush with respect to the fresh gases), by an adequate closure relation, and the resulting model is generally referred to as a Turbulent Flame velocity Closure (TFC) model [18]. The transport equation for the characteristic function of the burnt zone is called in this context the G-equation, its unknown being denoted by G [14]. Such a modelling is implemented in the in-house software P^2REMICS (for Partially PREMIxed Combustion Solver) developed, on the basis of the software components library CALIF^{3}S (for Components Adaptative Library For Fluid Flow Simulations, see [2]) at the French Institut de Radioprotection et de Sûreté Nucléaire (IRSN) for safety evaluation purposes; this is the context of the work presented in this paper.

Usually, TFC models apply to perfectly premixed flows (*i.e.* flows with constant initial composition), and the chemical state of the flow is governed by the value of G only: $G \in [0, 1]$, for $G \geq 0.5$, the mixture is supposed to be in its fresh (initial) state and $G < 0.5$ is supposed to correspond to the burnt state; in both cases, the composition of the gas is known (it is equal to the initial value in the fresh zones, and to the state resulting from a complete chemical reaction in the burnt zone).

However, for partially premixed turbulent flows (*i.e.* flows with non-constant initial composition), the situation is more complex, since the composition of the mixture can no more be deduced from the value of G. An extension for this situation, in the inviscid case, is proposed in [1]. The line followed to formulate this model is to write transport equations for the chemical species initially present in the flow, as if no chemical reaction occured, and then to compute the actual composition in the burnt zone (*i.e.* the part of the physical space where $G < 0.5$) as the chemical equilibrium composition, thus supposing an infinitely fast reaction. This model is referred to in the following as the *"asymptotic model"*, and is recalled in the first part of Sect. 2.

We propose here an alternate extension, which consists in keeping the classical reactive formulation of the chemical species mass balance, but evaluating the reaction term as a function of G: it is set to zero in the fresh zone ($G \geq 0.5$), and to a finite (but possibly large) value in the burnt zone ($G < 0.5$). This model is referred to as

the *"relaxed model"*; it is in fact more general, as it can be readily extended to cope with diffusion terms, while the "asymptotic model" cannot (to this purpose, a balance for the actual mass fractions is necessary). We then build a numerical scheme, based on a staggered discretization of the unknowns, for the solution of the relaxed model; this algorithm is of fractional step type, and employs a pressure correction technique for hydrodynamics. The balance energy solved by the scheme is the so-called (non conservative) sensible enthalpy balance, with corrective terms in order to ensure the weak consistency (in the Lax-Wendroff sense) of the scheme. It enjoys the same stability properties as the continuous model: positivity of the density and, thanks to the choice of the enthalpy balance, the internal energy, conservation of the total energy, chemical species mass fractions lying in the interval [0, 1]. In addition, it is shown to be in fact conservative: indeed, its solutions satisfy a discrete conservative total energy balance whose time and space discretization is non-standard, but weakly consistent with its continuous counterpart. This algorithm is an extension to the reactive case of the numerical scheme for compressible Navier-Stokes equations described and tested in [8].

As the reaction term gets stiffer, the relaxed model should boil down to the asymptotic one, for which a closed form of the solution of Riemann problems is available. Numerical tests are performed which show that this is indeed the case. In addition, we observe that the accuracy of the scheme (for this kind of application) is highly dependent on the numerical diffusion introduced by the scheme in the mass balance equation for the chemical species, comparing the results for three approximations of the convection operator in these equations: the standard upwind scheme, a MUSCL-like scheme introduced in [15] and a first order scheme designed to reduce diffusion proposed in [5].

The presentation is structured as follows. We first introduce the asymptotic and the relaxed models in Sect. 2. Then we give an overview of the content of this paper in Sect. 3, writing the scheme in the time semi-discrete setting and stating its stability and consistency property. The fully discrete setting is given in two steps, first describing the space discretization (Sect. 4) and then the scheme itself (Sect. 5). The conservativity of the scheme is shown in Sect. 6. Finally, numerical experiments are presented in Sect. 7.

2 The Physical Models

We begin with the description of the asymptotic model introduced in [1] and then turn to the relaxed model proposed in the present work.

The asymptotic model - For the sake of simplicity, only four chemical species are supposed to be present in the flow, namely the fuel (denoted by F), the oxydant (O), the product (P) of the reaction, and a neutral gas (N). A one-step irreversible total chemical reaction is considered, which is written:

$$\nu_F F + \nu_O O + N \rightarrow \nu_P P + N,$$

where ν_F, ν_O and ν_P are the molar stoichiometric coefficients of the reaction. We denote by $\mathcal{I}$ the set of the subscripts used to refer to the chemical species in the flow, so $\mathcal{I} = \{F, O, N, P\}$ and the set of mass fractions of the chemical species in the flow reads $\{y_i,\ i \in \mathcal{I}\}$ (*i.e.* $\{y_F,\ y_O,\ y_N,\ y_P\}$). We now define the auxiliary unknowns $\{\tilde{y}_i,\ i \in \mathcal{I}\}$ as the result of the (inert) transport by the flow of the initial state, which means that the $\{\tilde{y}_i,\ i \in \mathcal{I}\}$ are the solutions to the following system of equations:

$$\partial_t(\rho\tilde{y}_i) + \operatorname{div}(\rho\tilde{y}_i\boldsymbol{u}) = 0, \quad \tilde{y}_i(\boldsymbol{x}, 0) = y_{i,0}(\boldsymbol{x}), \quad \text{for } i \in \mathcal{I}, \tag{1}$$

where ρ stands for the fluid density, $\boldsymbol{u}$ for the velocity, and $y_{i,0}(\boldsymbol{x})$ is the initial mass fraction of the chemical species i in the flow. These equations are supposed to be posed over a bounded domain Ω of $\mathbb{R}^d$, $d \in \{1, 2, 3\}$ and a finite time interval $(0, T)$. The initial conditions are supposed to verify $\sum_{i\in\mathcal{I}} y_{i,0} = 1$ everywhere in Ω, and this property is assumed to be valid for any $t \in (0, T)$, which is equivalent with the mixture mass balance, given below. The characteristic function G is supposed to obey the following equation:

$$\partial_t(\rho G) + \operatorname{div}(\rho G\boldsymbol{u}) + \rho_u u_f |\nabla G| = 0, \tag{2}$$

associated to the initial conditions $G = 0$ at the location where the flame starts and $G = 1$ elsewhere. The quantity ρ_u is a constant density, which, from a physical point of view, stands for a characteristic value for the unburnt gases density. The chemical mass fractions are now computed as:

$$\left|\begin{array}{l} \text{if } G > 0.5,\ y_i = \tilde{y}_i \quad \text{for } i \in \mathcal{I}, \\ \text{if } G \le 0.5,\ y_F = \nu_F W_F \tilde{z}^+,\ y_O = \nu_O W_O \tilde{z}^-,\ y_N = \tilde{y}_N, \\ \qquad\qquad \text{with } \tilde{z} = \dfrac{1}{\nu_F W_F}\tilde{y}_F - \dfrac{1}{\nu_O W_O}\tilde{y}_O. \end{array}\right. \tag{3}$$

In these relation, $\tilde{z}^+$ and $\tilde{z}^-$ stand for the positive and negative part of $\tilde{z}$, respectively, *i.e.* $\tilde{z}^+ = \max(\tilde{z}, 0)$ and $\tilde{z}^- = -\min(\tilde{z}, 0)$, and, for $i \in \mathcal{I}$, W_i is the molar mass of the chemical species i. The physical meaning of Relation (3) is that the chemical reaction is supposed to be infinitely fast, and thus that the flow composition is stuck to the chemical equilibrium composition in the so-called burnt zone, which explains why the model is qualified as "asymptotic". The product mass fraction is given by $y_P = 1 - (y_F + y_O + y_N)$. The flow is governed by the Euler equations:

$$\partial_t\rho + \operatorname{div}(\rho\boldsymbol{u}) = 0, \tag{4a}$$

$$\partial_t(\rho u_i) + \operatorname{div}(\rho u_i\boldsymbol{u}) + \partial_i p = 0, \quad i = 1, d, \tag{4b}$$

$$\partial_t(\rho E) + \operatorname{div}(\rho E\boldsymbol{u}) + \operatorname{div}(p\boldsymbol{u}) = 0, \tag{4c}$$

$$p = (\gamma - 1)\,\rho e_s, \qquad E = \frac{1}{2}|\boldsymbol{u}|^2 + e, \quad e = e_s + \sum_{i\in\mathcal{I}} y_i \Delta h^0_{f,i}, \tag{4d}$$

where p stands for the pressure, E for the total energy, e for the internal energy, e_s for the so-called sensible internal energy and, for $i \in \mathcal{I}$, $\Delta h^0_{f,i}$ is the formation enthalpy of the chemical species i. The equation of state (4d) supposes that the fluid is a perfect mixture of ideal gases, with the same iso-pressure to iso-volume specific heat ratio $\gamma > 1$. This set of equations is complemented by homogeneous Neumann boundary conditions for the velocity:

$$\boldsymbol{u} \cdot \boldsymbol{n} = 0 \quad \text{a.e. on } \partial\Omega, \tag{5}$$

where $\partial\Omega$ stands for the boundary of Ω and $\boldsymbol{n}$ its outward normal vector.

The "relaxed" model – This model retains the original form of the governing equations for reactive flows: a transport/reaction equation is written for each of the chemical species mass fractions; the value of G controls the reaction rate $\dot{\omega}$, which is set to zero when $G \geq 0.5$, and takes non-zero (and possibly large) values otherwise. The unknowns $\{y_i,\ i \in \mathcal{I}\}$ are thus now solution to the following balance equations:

$$\partial_t(\rho y_i) + \mathrm{div}(\rho y_i \boldsymbol{u}) = \dot{\omega}_i, \quad \tilde{y}_i(\boldsymbol{x}, 0) = y_{i,0}(\boldsymbol{x}) \qquad \text{for } i \in \mathcal{I}, \tag{6}$$

where the reactive term $\dot{\omega}_i$ is given by:

$$\dot{\omega}_i = \frac{1}{\varepsilon}\, \zeta_i\, \nu_i W_i\, \dot{\omega}, \ \text{ with } \dot{\omega} = \eta(y_F, y_O)\ (G - 0.5)^-$$
$$\text{and } \eta(y_F, y_O) = \min\left(\frac{y_F}{\nu_F W_F}, \frac{y_O}{\nu_O W_O}\right), \tag{7}$$

with $\zeta_F = \zeta_O = -1$, $\zeta_P = 1$ and $\zeta_N = 0$. Note that, since $\nu_F W_F + \nu_O W_0 = \nu_P W_P$, we have $\sum_{i\in\mathcal{I}} \dot{\omega}_i = 0$, which, summing on $i \in \mathcal{I}$ the species mass balances, allows to recover the equivalence between the mass balance and the fact that $\sum_{i\in\mathcal{I}} y_i = 1$. The factor $\eta(y_F, y_O)$ is a cut-off function, which prevents the chemical species mass fractions from taking negative values (and, consequently, values greater than 1, since their sum is equal to 1).

The rest of the model is left unchanged.

3 General Description of the Scheme and Main Results

Time Semi-discrete Algorithm

Instead of the total energy balance equation, the scheme solves a balance equation for the sensible enthalpy $h_s = e_s + p/\rho$, which is formally derived as follows. The first step is to establish the kinetic energy balance formally and subtract from (4c) to obtain a balance equation for the internal energy. Thanks to the mass balance equation, for any regular function ψ

$$\partial_t(\rho\psi) + \mathrm{div}(\rho\psi\boldsymbol{u}) = \rho\partial_t\psi + \rho\boldsymbol{u}\cdot\boldsymbol{\nabla}\psi.$$

Using twice this identity and then the momentum balance equation, we have for $1 \le i \le d$:

$$\begin{aligned}\frac{1}{2}\partial_t(\rho u_i^2) + \frac{1}{2}\mathrm{div}(\rho u_i^2\,\boldsymbol{u}) &= \rho\, u_i\partial_t u_i + \rho u_i\boldsymbol{u}\cdot\boldsymbol{\nabla}u_i \\ &= u_i\big[\partial_t(\rho u_i) + \mathrm{div}(\rho u_i\boldsymbol{u})\big] = -u_i\,\partial_i\, p,\end{aligned}$$

and, summing for $i = 1$ to d, we obtain the kinetic energy balance:

$$\frac{1}{2}\partial_t(\rho|\boldsymbol{u}|^2) + \frac{1}{2}\mathrm{div}(\rho|\boldsymbol{u}|^2\boldsymbol{u}) = \boldsymbol{u}\cdot\big[\partial_t(\rho\boldsymbol{u}) + \mathrm{div}(\rho\boldsymbol{u}\otimes\boldsymbol{u})\big] = -\boldsymbol{u}\cdot\boldsymbol{\nabla}p.$$

Substituting the expression of the total energy in (4c), yields

$$\partial_t(\rho e) + \mathrm{div}(\rho e\boldsymbol{u}) + \frac{1}{2}\partial_t(\rho|\boldsymbol{u}|^2) + \frac{1}{2}\mathrm{div}(\rho|\boldsymbol{u}|^2) + \boldsymbol{u}\cdot\boldsymbol{\nabla}p + p\mathrm{div}(\boldsymbol{u}) = 0,$$

which, using the kinetic energy balance, gives the total internal energy balance:

$$\partial_t(\rho e) + \mathrm{div}(\rho e\boldsymbol{u}) + p\mathrm{div}(\boldsymbol{u}) = 0. \tag{8}$$

Using the linearity of the mass balance of the chemical species i, for any $i \in \mathcal{I}$, we derive the reactive energy balance:

$$\partial_t\big[\rho\big(\sum_{i\in\mathcal{I}}\Delta h^0_{f,i}\,y_i\big)\big] + \mathrm{div}\big[\rho\big(\sum_{i\in\mathcal{I}}\Delta h^0_{f,i}\,y_i\big)\boldsymbol{u}\big] = \sum_{i\in\mathcal{I}}\Delta h^0_{f,i}\dot{\omega}_i = -\dot{\omega}_\theta. \tag{9}$$

Subtracting (9) from (8) yields the sensible internal energy balance:

$$\partial_t(\rho e_s) + \mathrm{div}(\rho e_s\boldsymbol{u}) + p\mathrm{div}(\boldsymbol{u}) = \dot{\omega}_\theta. \tag{10}$$

Finally, using the relation between the sensible energy and the sensible enthalpy, we obtain the sensible enthalpy balance:

$$\partial_t(\rho h_s) + \mathrm{div}(\rho h_s\boldsymbol{u}) - \partial_t p - \boldsymbol{u}\cdot\boldsymbol{\nabla}p = \dot{\omega}_\theta. \tag{11}$$

The numerical resolution of the mathematical model is realized by a fractional step algorithm, which implements a pressure correction technique for hydrodynamics in order to separate the resolution of the momentum balance from the other equations of the Euler system. Supposing that the time interval $(0, T)$ is split in N sub-intervals, of constant length $\delta t = T/N$, the semi-discrete algorithm is given by:

Reactive step:

$$G^{n+1} : \frac{1}{\delta t}(\rho^n G^{n+1} - \rho^{n-1} G^n) + \operatorname{div}(\rho^n G^k \boldsymbol{u}^n) + \rho_u u_f \, |\nabla G^k| = 0, \tag{12a}$$

$$Y_N^{n+1} : \frac{1}{\delta t}(\rho^n y_N^{n+1} - \rho^{n-1} y_N^n) + \operatorname{div}(\rho^n y_N^k \boldsymbol{u}^n) = 0. \tag{12b}$$

$$z^{n+1} : \frac{1}{\delta t}(\rho^n z^{n+1} - \rho^{n-1} z^n) + \operatorname{div}(\rho^n z^k \boldsymbol{u}^n) = 0. \tag{12c}$$

$$Y_F^{n+1} : \frac{1}{\delta t}(\rho^n y_F^{n+1} - \rho^{n-1} y_F^n) + \operatorname{div}(\rho^n y_F^k \boldsymbol{u}^n) = -\frac{1}{\varepsilon} \nu_F W_F \, \dot{\omega}(y_F^{n+1}, z^{n+1}), \tag{12d}$$

$$Y_P^{n+1} : y_F^{n+1} + y_O^{n+1} + y_N^{n+1} + y_P^{n+1} = 1. \tag{12e}$$

Euler step:

$$\tilde{\boldsymbol{u}}^{n+1} : \frac{1}{\delta t}(\rho^n \tilde{u}_i^{n+1} - \rho^{n-1} u_i^n) + \operatorname{div}(\rho^n \tilde{u}_i^{n+1} \boldsymbol{u}^n) + \left(\frac{\rho^n}{\rho^{n-1}}\right)^{1/2} \partial_i p^n = 0, \quad i = 1, \ldots, d, \tag{12f}$$

$$\begin{matrix} \boldsymbol{u}^{n+1} \\ \rho^{n+1} \\ h_s^{n+1} \\ p^{n+1} \end{matrix} : \left| \begin{array}{l} \dfrac{1}{\delta t}\rho^n (u_i^{n+1} - \tilde{u}_i^{n+1}) + \partial_i p^{n+1} - \left(\dfrac{\rho^n}{\rho^{n-1}}\right)^{1/2} \partial_i p^n = 0, \ i = 1, \ldots, d, \\ \dfrac{1}{\delta t}(\rho^{n+1} - \rho^n) + \operatorname{div}(\rho^{n+1} \boldsymbol{u}^{n+1}) = 0, \\ \dfrac{1}{\delta t}(\rho^{n+1} h_s^{n+1} - \rho^n h_s^n) + \operatorname{div}(\rho^{n+1} h_s^{n+1} \boldsymbol{u}^{n+1}) - \dfrac{1}{\delta t}(p^{n+1} - p^n) - u^{n+1} \cdot \nabla p^{n+1} = \dot{\omega}_\theta^{n+1} + S^{n+1}, \\ p^{n+1} = \dfrac{\gamma - 1}{\gamma} \rho^{n+1} h_s^{n+1}. \end{array} \right. \tag{12g}$$

Equations (12a)–(12g) are solved successively, and the unknown for each equation is specified before each equation. In the convection term of the equations of the reactive step, the index k may take the value n (explicit scheme) or $n+1$ (implicit scheme). The unknown z is an affine combination of y_F and y_O, defined so that the reactive term cancels:

$$z = \frac{1}{\nu_F W_F} y_F - \frac{1}{\nu_O W_O} y_O. \tag{13}$$

Thus the value of y_O^{n+1} is deduced from y_F^{n+1} and z^{n+1}, which allows to express $\dot{\omega}$ in (12d) as a function of y_F^{n+1} and z^{n+1}, instead of y_F^{n+1} and y_O^{n+1} as suggested by Relation (7). In addition, we have:

$$\eta(y_F^{n+1}, y_O^{n+1}) = \min\left(\frac{y_F^{n+1}}{\nu_F W_F}, \frac{y_O^{n+1}}{\nu_O W_O}\right)$$

$$= \begin{cases} \dfrac{1}{\nu_F W_F}\, y_F^{n+1} & \text{if } z^{n+1} \le 0, \\ \dfrac{1}{\nu_O W_O}\, y_O^{n+1} = \dfrac{1}{\nu_F W_F}\, y_F^{n+1} - z^{n+1} & \text{otherwise.} \end{cases}$$

Hence, because of the specific form of the function η, the right hand side of (12d) boils down to an affine term, even if η vanishes when y_F or y_O vanishes, and the scheme is fully implicit in time with respect to the reaction term. This is the motivation for the choice of the form of η. It is fundamental to remark that Eqs. (12b)–(12e) are equivalent to the following system:

$$\frac{1}{\delta t}(\rho^n y_i^{n+1} - \rho^{n-1} y_i^n) + \operatorname{div}(\rho^n y_i^k \boldsymbol{u}^n) = \frac{1}{\varepsilon} \zeta_i \nu_i W_i\, \dot{\omega}(y_F^{n+1}, y_O^{n+1}),\ i \in \mathcal{I}, \tag{14}$$

where we recall that $\zeta_F = \zeta_O = -1$, $\zeta_P = 1$ and $\zeta_N = 0$. Indeed, dividing the fuel mass balance equation (12d) by $\nu_F W_F$, substracting Eq. (12c) and finally multiplying by $\nu_O W_O$ yields the desired mass balance equation for the oxydant chemical species. Finally, we suppose that the product mass balance holds:

$$\frac{1}{\delta t}(\rho^n y_P^{n+1} - \rho^{n-1} y_P^n) + \operatorname{div}(\rho^n y_P^k \boldsymbol{u}^n) = \frac{1}{\varepsilon} \nu_P W_P\, \dot{\omega}(y_F^{n+1}, y_O^{n+1}). \tag{15}$$

Since the sum of the chemical reaction terms vanishes, we have for $\Sigma = y_F + y_O + y_P + y_N$, summing all the chemical species mass balances,

$$\frac{1}{\delta t}(\rho^n \Sigma^{n+1} - \rho^{n-1} \Sigma^n) + \operatorname{div}(\rho^n \Sigma^k \boldsymbol{u}^n) = 0, \tag{16}$$

and this equation may equivalently replace the product mass balance equation (15). Thanks to the mixture balance, we see that, provided that Σ^n satisfies $\Sigma^n = 1$ everywhere in Ω, the solution to Eq. (16) is $\Sigma^{n+1} = 1$ everywhere in Ω. Since the initialization yields $\Sigma^0 = 1$, this last equality is indeed true, and (15) is equivalent to (12e). Finally, note that, when the chemical step is performed, the mass balance at step $n+1$ is not yet solved; hence the (unusual) backward time shift for the densities and for the mass fluxes in the equations of this step.

Equations (12f)–(12g) implement a pressure correction technique, where the correction step couples the velocity correction equation, the mass balance and the sensible enthalpy balance. This coupling ensures that the pressure and velocity are kept constant through the contact discontinuity associated to compositional non-reactive Euler equations (precisely speaking, the usual contact discontinuity, already present in 1D equations, but not slip lines); for this property to hold, it is necessary that all chemical species share the same heat capacity ratio γ. The term S_K^{n+1} in the sensible enthalpy balance equation is a corrective term which is necessary for consistency; schematically speaking, it compensates the numerical dissipation which appears in a discrete kinetic energy balance that is obtained from the discrete momentum balance.

Its expression is given in Sect. 5, and its derivation is explained in Sect. 6, where the conservativity of the scheme is discussed.

Space Discretization

The space dicretization is performed by a finite volume technique, using a staggered arrangement of the unknowns (the scalar variables are approximated at the cell centers and the velocity components at the face centers), using either a MAC scheme (for structured discretizations) or the degrees of freedom of low-order non-conforming finite elements: Crouzeix-Raviart [4] for simplicial cells and Rannacher-Turek [17] for quadrangles ($d = 2$) or hexahedra ($d = 3$). For the Euler equations (*i.e.* steps (12f)–(12g)), upwinding is performed by building positivity-preserving convection operators, in the spirit of the so-called Flux-Splitting methods, and only first-order upwinding is implemented. The pressure gradient is built as the transpose (with respect to the L^2 inner product) of the natural velocity divergence operator. For the balance equations for the other scalar unknowns, the time discretization is implicit when first-order upwinding is used in the convection operator (in other words, $k = n + 1$ in (12a)–(12d)) or explicit ($k = n$ in (12a)–(12d)) when a higher order (of MUSCL type, *cf.* Appendix 8) flux or an anti-diffusive flux (*cf.* Appendix 9) is used.

Properties of the Scheme

First, the positivity of the density is ensured by construction of the discrete mass balance equation, *i.e.* by the use of a first order upwind scheme. In addition, the physical bounds of the mass fractions are preserved thanks to the following (rather standard) arguments: first, building a discrete convection operator which vanishes when the convected unknown is constant thanks to the discrete mass balance equation ensures a positivity-preservation property [13], under a CFL condition if an explicit time approximation is used; second, the discretization of the chemical reaction rate ensures either that it vanishes when the unknown of the equation vanishes (for y_F and y_O), or that it is non-negative (for y_P). Consequently, mass fractions are non-negative and, since their sum is equal to 1 (see above), they are also bounded by 1.

The positivity of the sensible energy stems from two essential arguments: first, a discrete analog of the internal energy equation (8) may be obtained from the discrete sensible enthalpy balance, by mimicking the continuous computation; second, this discrete relation may be shown to have only positive solutions, once again thanks to the consistency of the discrete convection operator and the mass balance. This holds provided that the equation is exothermic ($\dot{\omega}_\theta \geq 0$) and thanks to the non-negativity of S^{n+1} (see below).

In order to calculate correct shocks, it is crucial for the scheme to be consistent with the following weak formulation of the problem:

$$
\begin{aligned}
&\forall \phi \in C_c^\infty(\Omega \times [0,T)\big),\\
&\int_0^T\!\!\int_\Omega \big[\rho \partial_t \phi + \rho \boldsymbol{u}\cdot\nabla\phi\big]\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega \rho_0(\boldsymbol{x})\phi(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} = 0,\\
&\int_0^T\!\!\int_\Omega \big[\rho u_i \partial_t \phi + (\rho \boldsymbol{u} u_i)\cdot\nabla\phi + p\partial_i\phi\big]\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\\
&\qquad\qquad + \int_\Omega \rho_0(\boldsymbol{x})(u_i)_0(\boldsymbol{x})\phi(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} = 0,\ 1\le i\le d,\\
&\int_0^T\!\!\int_\Omega \big[\rho E \partial_t \phi + (\rho E + p)\boldsymbol{u}\cdot\nabla\phi\big]\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_\Omega \rho_0(\boldsymbol{x})E_0(\boldsymbol{x})\phi(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} = 0,\\
&\int_0^T\!\!\int_\Omega \big[\rho y_i \partial_t \phi + \rho y_i \boldsymbol{u}\cdot\nabla\phi\big]\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_0^T\!\!\int_\Omega \rho_0(\boldsymbol{x})y_{i,0}(\boldsymbol{x})\phi(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} =\\
&\qquad\qquad - \int_0^T\!\!\int_\Omega \dot\omega_i\phi\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t,\ 1\le i\le d,\\
&p = (\gamma-1)\rho e_s.
\end{aligned}
\tag{17}
$$

Remark that this system features the total energy balance equation and not the sensible enthalpy balance equation, which is actually solved here. However, we show in Sect. 6 that the solutions of the scheme satisfy a discrete total energy balance, with a time and space dicretization which is unusual but allows however to prove the consistency in the Lax-Wendroff sense. Finally, the integral of the total energy over the domain is conserved, which yields a stability result for the scheme (irrespectively of the time and space step, for this relation; recall however that the overall stability of the scheme needs a CFL condition if an explicit version of the convection operator for chemical species is used).

4 Meshes and Unknowns

Let the computational domain Ω be an open polygonal subset of $\mathbb{R}^d$, $1 \le d \le 3$, with boundary $\partial\Omega$ and let $\mathcal{M}$ be a decomposition of Ω, supposed to be regular in the usual sense of the finite element literature (*e.g.* [3]). The cells may be:

- for a general domain Ω, either convex quadrilaterals ($d = 2$) or hexahedra ($d = 3$) or simplices, both type of cells being possibly combined in a same mesh for two-dimensional problems,
- for a domain whose boundaries are hyperplanes normal to a coordinate axis, rectangles ($d = 2$) or rectangular parallelepipeds ($d = 3$) (and whose faces, of course, are then also necessarily normal to a coordinate axis).

By $\mathcal{E}$ and $\mathcal{E}(K)$ we denote the set of all $(d-1)$-faces σ of the mesh and of the element $K \in \mathcal{M}$ respectively. The set of faces included in the boundary of Ω is denoted by $\mathcal{E}_{\text{ext}}$ and the set of internal edges (*i.e.* $\mathcal{E} \setminus \mathcal{E}_{\text{ext}}$) is denoted by $\mathcal{E}_{\text{int}}$; a face $\sigma \in \mathcal{E}_{\text{int}}$ separating the cells K and L is denoted by $\sigma = K|L$. The outward normal vector to a face σ of K is denoted by $\boldsymbol{n}_{K,\sigma}$. For $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}$, we denote by

$|K|$ the measure of K and by $|\sigma|$ the $(d-1)$-measure of the face σ. The size of the mesh is denoted by h:

$$h = \max\{\mathrm{diam}(K),\ K \in \mathcal{M}\}.$$

For $1 \leq i \leq d$, we denote by $\mathcal{E}^{(i)} \subset \mathcal{E}$ and $\mathcal{E}_{\mathrm{ext}}^{(i)} \subset \mathcal{E}_{\mathrm{ext}}$ the subset of the faces of $\mathcal{E}$ and $\mathcal{E}_{\mathrm{ext}}$ respectively which are perpendicular to the i^{th} unit vector of the canonical basis of $\mathbb{R}^d$.

The space discretization is staggered, using either the Marker-And Cell (MAC) scheme [9, 10], or nonconforming low-order finite element approximations, namely the Rannacher and Turek (RT) element [17] for quadrilateral or hexahedric meshes, or the lowest degree Crouzeix-Raviart (CR) element [4] for simplicial meshes.

For all these space discretizations, the degrees of freedom for the pressure, the density, the enthalpy, the mixture, fuel and neutral gas mass fractions and the flame indicator are associated to the cells of the mesh $\mathcal{M}$ and are denoted by:

$$\{p_K,\ \rho_K,\ h_K,\ y_{F,K},\ y_{N,K},\ z_K,\ G_K,\ K \in \mathcal{M}\}.$$

Let us then turn to the degrees of freedom for the velocity (*i.e.* the discrete velocity unknowns).

– **Rannacher-Turek** or **Crouzeix-Raviart** discretizations – The degrees of freedom for the velocity components are located at the center of the faces of the mesh, and we choose the version of the element where they represent the average of the velocity through a face. The set of degrees of freedom reads:

$$\{\boldsymbol{u}_\sigma,\ \sigma \in \mathcal{E}\},\ \text{of components } \{u_{\sigma,i},\ \sigma \in \mathcal{E},\ 1 \leq i \leq d\}.$$

– **MAC** discretization – The degrees of freedom for the i^{th} component of the velocity are defined at the centre of the faces of $\mathcal{E}^{(i)}$, so the whole set of discrete velocity unknowns reads:

$$\{u_{\sigma,i},\ \sigma \in \mathcal{E}^{(i)},\ 1 \leq i \leq d\}.$$

For the definition of the schemes, we need a dual mesh which is defined as follows.

– **Rannacher-Turek** or **Crouzeix-Raviart** discretizations – For the RT or CR discretizations, the dual mesh is the same for all the velocity components. When $K \in \mathcal{M}$ is a simplex, a rectangle or a rectangular cuboid, for $\sigma \in \mathcal{E}(K)$, we define $D_{K,\sigma}$ as the cone with basis σ and with vertex the mass center of K (see Fig. 1). We thus obtain a partition of K in m sub-volumes, where m is the number of faces of the mesh, each sub-volume having the same measure $|D_{K,\sigma}| = |K|/m$. We extend this definition to general quadrangles and hexahedra, by supposing that we have built a partition still of equal-volume sub-cells, and with the same connectivities; note that this is of course always possible, but that such a volume $D_{K,\sigma}$ may be no longer a cone; indeed, if K is far from a parallelogram, it may not be possible to build a cone having σ as basis, the opposite vertex lying in K and a volume

equal to $|K|/m$ (note that these dual cells do not need to be constructed in the implementation of the scheme, only their volume is needed). The volume $D_{K,\sigma}$ is referred to as the half-diamond cell associated to K and σ.
For $\sigma \in \mathcal{E}_{\text{int}}$, $\sigma = K|L$, we now define the diamond cell D_σ associated to σ by $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$; for an external face $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}(K)$, D_σ is just the same volume as $D_{K,\sigma}$.

– **MAC** discretization – For the MAC scheme, the dual mesh depends on the component of the velocity. For each component, the MAC dual mesh only differs from the RT or CR dual mesh by the choice of the half-diamond cell, which, for $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}(K)$, is now the rectangle or rectangular parallelepiped of basis σ and of measure $|D_{K,\sigma}| = |K|/2$.

We denote by $|D_\sigma|$ the measure of the dual cell D_σ, and by $\varepsilon = D_\sigma|D_{\sigma'}$ the dual face separating two diamond cells D_σ and $D_{\sigma'}$.

In order to be able to write a unique expression of the discrete equations for both MAC and CR/RT schemes, we introduce the set of faces $\mathcal{E}_{\mathcal{S}}^{(i)}$ associated with the degrees of freedom of each component of the velocity ($\mathcal{S}$ stands for "scheme"):

$$\mathcal{E}_{\mathcal{S}}^{(i)} = \begin{cases} \mathcal{E}^{(i)} \setminus \mathcal{E}_{\text{ext}}^{(i)} \text{ for the MAC scheme,} \\ \mathcal{E} \setminus \mathcal{E}_{\text{ext}}^{(i)} \text{ for the CR or RT schemes.} \end{cases}$$

Similarly, we unify the notation for the set of dual faces for both schemes by defining:

$$\tilde{\mathcal{E}}_{\mathcal{S}}^{(i)} = \begin{cases} \tilde{\mathcal{E}}^{(i)} \setminus \tilde{\mathcal{E}}_{\text{ext}}^{(i)} \text{ for the MAC scheme,} \\ \tilde{\mathcal{E}} \setminus \tilde{\mathcal{E}}_{\text{ext}}^{(i)} \text{ for the CR or RT schemes,} \end{cases}$$

where the symbol $\tilde{}$ refers to the dual mesh; for instance, $\tilde{\mathcal{E}}^{(i)}$ is thus the set of faces of the dual mesh associated with the i^{th} component of the velocity, and $\tilde{\mathcal{E}}_{\text{ext}}^{(i)}$ stands for the subset of these dual faces included in the boundary. Note that, for the MAC scheme, the faces of $\tilde{\mathcal{E}}^{(i)}$ are perpendicular to a unit vector of the canonical basis of $\mathbb{R}^d$, but not necessarily to the i^{th} one.

5 The Scheme

In this section, we give the fully discrete form of the scheme. Even if it corresponds to the reverse order with respect to the semi-discrete scheme given in (12), we begin with the hydrodynamics (Sect. 5.1) and then turn to the mass balance step for chemical species and the transport of the characteristic function for the burnt zone (Sect. 5.2). This choice is due to the fact that the definition of the convection operators for scalar variables necessitates to introduce the discretization of the mixture mass balance equation first.

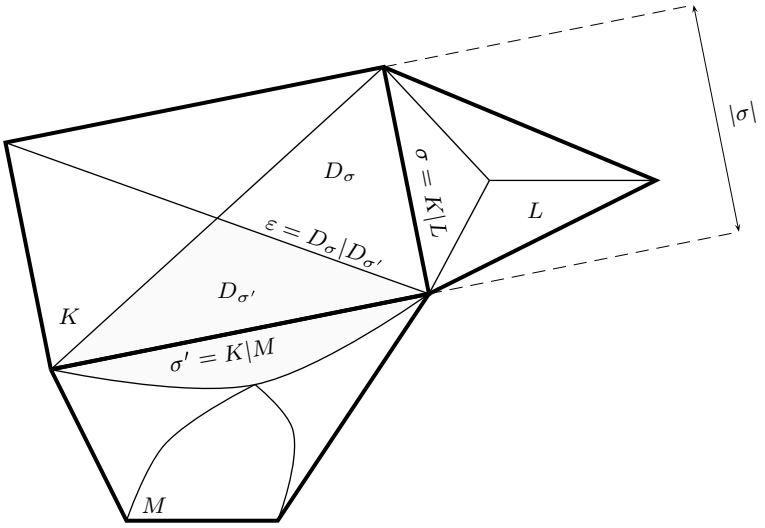


Fig. 1 Primal and dual meshes for the Rannacher-Turek and Crouzeix-Raviart elements.

5.1 Euler Step

For $0 \leq n < N$, the step $n+1$ of the algorithm for the resolution of the Euler equations reads:

Pressure gradient scaling step – Solve for $(\widetilde{\nabla p})^{n+1}$:

$$\forall \sigma \in \mathcal{E}, \qquad (\widetilde{\nabla p})_\sigma^{n+1} = \Big(\frac{\rho_{D_\sigma}^n}{\rho_{D_\sigma}^{n-1}}\Big)^{1/2} (\nabla p)_\sigma^n. \tag{18a}$$

Prediction step – Solve for $\tilde{\boldsymbol{u}}^{n+1}$:

For $1 \leq i \leq d$, $\forall \sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}$,

$$\frac{1}{\delta t}(\rho_{D_\sigma}^n \tilde{u}_{\sigma,i}^{n+1} - \rho_{D_\sigma}^{n-1} u_{\sigma,i}^n) + \mathrm{div}_\sigma(\rho^n \tilde{u}_i^{n+1} \boldsymbol{u}^n) + (\widetilde{\nabla p})_{\sigma,i}^{n+1} = 0. \tag{18b}$$

Correction step – Solve for ρ^{n+1}, p^{n+1} and $\boldsymbol{u}^{n+1}$:

For $1 \leq i \leq d$, $\forall \sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}$,

$$\frac{1}{\delta t}\, \rho_{D_\sigma}^n\, (u_{\sigma,i}^{n+1} - \tilde{u}_{\sigma,i}^{n+1}) + (\nabla p)_{\sigma,i}^{n+1} - (\widetilde{\nabla p})_{\sigma,i}^{n+1} = 0, \tag{18c}$$

$$\forall K \in \mathcal{M},\ \frac{1}{\delta t}(\rho_K^{n+1} - \rho_K^n) + \mathrm{div}_K(\rho \boldsymbol{u})^{n+1} = 0, \tag{18d}$$

$$\forall K \in \mathcal{M},\ \frac{1}{\delta t}\left[\rho_K^{n+1}\,(h_s)_K^{n+1} - \rho_K^n\,(h_s)_K^n\right] + \mathrm{div}_K(\rho h_s \boldsymbol{u})^{n+1} - \frac{1}{\delta t}\,(p_K^{n+1} - p_K^n) - \big(u \cdot \nabla p\big)_K^{n+1} = (\dot{\omega}_\theta)_K^{n+1} + S_K^{n+1}, \tag{18e}$$

$$\forall K \in \mathcal{M}, \qquad p_K^{n+1} = \frac{\gamma - 1}{\gamma} (h_s)_K^{n+1} \rho_K^{n+1}. \tag{18f}$$

The initial approximations for ρ^{-1}, h_s^0 and $\boldsymbol{u}^0$ are given by the mean values of the initial conditions over the primal and dual cells:

$$\forall K \in \mathcal{M}, \quad \rho_K^{-1} = \frac{1}{|K|} \int_K \rho_0(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \quad \text{and} \quad (h_s)_K^0 = \frac{1}{|K|} \int_K (h_s)_0(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x},$$
$$\forall \sigma \in \mathcal{E}_S^{(i)}, \ 1 \le i \le d, \quad u_{\sigma,i}^0 = \frac{1}{|D_\sigma|} \int_{D\sigma} (\boldsymbol{u}_0(\boldsymbol{x}))_i \, \mathrm{d}\boldsymbol{x}.$$

Then, ρ^0 is computed by the mass balance equation (18d) and p^0 is computed by the equation of state (18f).

We now define each of the discrete operators featured in System (18).

Mass Balance Equation. Equation (18d) is a finite volume discretisation of the mass balance (4a) over the primal mesh. For a discrete density field ρ and a discrete velocity field $\boldsymbol{u}$, the discrete divergence is defined by:

$$\mathrm{div}_K(\rho \boldsymbol{u}) = \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}, \quad F_{K,\sigma} = |\sigma| \, \rho_\sigma u_{K,\sigma},$$

where $u_{K,\sigma}$ is an approximation of the normal velocity to the face σ outward K. The definition of this latter quantity depends on the discretization: in the MAC case, $u_{K,\sigma} = u_{\sigma,i} \, \boldsymbol{e}^{(i)} \cdot \boldsymbol{n}_{K,\sigma}$ for a face σ of K perpendicular to $\boldsymbol{e}^{(i)}$, with $\boldsymbol{e}^{(i)}$ the i-th vector of the orthonormal basis of $\mathbb{R}^d$, and, in the CR and RT cases, $u_{K,\sigma} = \boldsymbol{u}_\sigma \cdot \boldsymbol{n}_{K,\sigma}$ for any face σ of K. The density at the face $\sigma = K|L$ is approximated by the upwind technique, so $\rho_\sigma = \rho_K$ if $u_{K,\sigma} \ge 0$ and $\rho_\sigma = \rho_L$ otherwise. Since we assume that the normal velocity vanishes on the boundary faces, the definition is complete.

Convection Operators Associated to the Primal Mesh

We may now give the general form of the discrete convection operator of any discrete field z defined on the primal cell:

$$\mathrm{div}_K(\rho z \boldsymbol{u}) = \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma} \, z_\sigma, \tag{19}$$

where z_σ is an approximation of the unknown z at the face σ.

Momentum Balance Equation and Pressure Gradient Scaling

We now turn to the discrete momentum balance (18b). For the MAC discretization, but also for the RT and CR discretizations, the time derivative and convection terms are approximated in (18b) by a finite volume technique over the dual cells, so the convection term reads:

$$\mathrm{div}_\sigma(\rho\tilde{u}_i\boldsymbol{u}) = \mathrm{div}_\sigma\big(\tilde{u}_i(\rho\boldsymbol{u})\big) = \frac{1}{|D_\sigma|}\sum_{\varepsilon\in\tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\varepsilon}\tilde{u}_{\varepsilon,i},$$

where $F_{\sigma,\varepsilon}$ stands for a mass flux through the dual face ε, and $\tilde{u}_{\varepsilon,i}$ is a centered approximation of the i^{th} component of the velocity $\tilde{\boldsymbol{u}}$ on ε. The density in the dual cell ρ_{D_σ} is obtained by a weighted average of the density in the neighbour cells: $|D_\sigma|\,\rho_{D_\sigma} = |D_{K,\sigma}|\,\rho_K + |D_{L,\sigma}|\,\rho_L$ for $\sigma = K|L \in \mathcal{E}_{\mathrm{int}}$, and $\rho_{D_\sigma} = \rho_K$ for an external face of a cell K. The mass fluxes $(F_{\sigma,\varepsilon})_{\varepsilon\in\mathcal{E}(D_\sigma)}$ are evaluated as linear combinations, with constant coefficients, of the primal mass fluxes at the neighbouring faces, in such a way that the following discrete mass balance over the dual cells is implied by the discrete mass balance (18d):

$$\forall\sigma\in\mathcal{E}\text{ and }n\in\mathbb{N},\qquad \frac{|D_\sigma|}{\delta t}\left(\rho_{D_\sigma}^{n+1}-\rho_{D_\sigma}^{n}\right)+\sum_{\varepsilon\in\mathcal{E}(D_\sigma)}F_{\sigma,\varepsilon}^{n+1}=0. \tag{20}$$

This relation is critical to derive a discrete kinetic energy balance (see Sect. 6 below). The computation of the dual mass fluxes is such that the flux through a dual face lying on the boundary, which is then also a primal face, is the same as the primal flux, that is zero. For the expression of these fluxes, we refer to [6, 11, 12]. Since the mass balance is not yet solved at the velocity prediction stage, they have to be built from the mass balance at the previous time step: hence the backward time shift for the densities in the time-derivative term.

The term $(\nabla p)_{\sigma,i}$ stands for the i-th component of the discrete pressure gradient at the face σ. This gradient operator is built as the transpose of the discrete operator for the divergence of the velocity, *i.e.* in such a way that the following duality relation with respect to the L^2 inner product holds:

$$\sum_{K\in\mathcal{M}}|K|\,p_K\,\mathrm{div}_K(\boldsymbol{u})+\sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}_{\mathcal{S}}^{(i)}}|D_\sigma|u_{\sigma,i}(\nabla p)_{\sigma,i}=0.$$

This leads to the following expression:

$$\forall\sigma=K|L\in\mathcal{E}_{\mathrm{int}},\qquad (\nabla p)_{\sigma,i}=\frac{|\sigma|}{|D_\sigma|}(p_L-p_K)\,\boldsymbol{n}_{K,\sigma}\cdot\boldsymbol{e}^{(i)}.$$

The scaling of the pressure gradient (18a) is necessary for the solution to the scheme to satisfy a local discrete (finite volume) kinetic energy balance [8, Lemma 4.1].

Sensible enthalpy equation The convection term for the sensible enthalpy takes the form (19), with an implicit and upwind (with respect to the mass flux $F_{K,\sigma}$) approximation of the unknown at the face. In addition, this equation is discretized in such a way that the present enthalpy formulation is strictly equivalent to the internal energy formulation of the energy balance equation used in [8]. Consequently, the term $-\big(u\cdot\nabla p\big)_K$ reads:

$$-\big(u\cdot\nabla p\big)_K=\frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)}|\sigma|\,u_{K,\sigma}\,(p_K-p_\sigma),$$

where p_σ is the upwind approximation of p at the face σ with respect to $u_{K,\sigma}$. The reaction heat, $(\dot\omega_\theta)_K$, is written in the following way:

$$(\dot\omega_\theta)_K=-\sum_{i=1}^{N_s}\Delta h^0_{f,i}(\dot\omega_i)_K=\big(\nu_F W_F\Delta h^0_{f,F}+\nu_O W_O\Delta h^0_{f,O}-\nu_P W_P\Delta h^0_{f,P}\big)\,\dot\omega_K.$$

The definition of $\dot\omega_K$ is given in Sect. 5.2, and the definition of the corrective term S_K^{n+1} is given in Sect. 6 (see Eq. (31) and Remark 3 below).

5.2 Chemistry Step

For $0\le n<N$, the step $n+1$ for the solution of the transport of the characteristic function of the burnt zone and the chemical species mass balance equations reads:

Computation of the burnt zone characteristic function – Solve for G^{n+1} :

$$\forall K\in\mathcal{M},\quad \frac{1}{\delta t}\,(\rho_K^n G_K^{n+1}-\rho_K^{n-1}G_K^n)+\mathrm{div}_K(\rho^n G^k\boldsymbol{u}^n)+(\rho_u u_f\,|\nabla G^k|)_K=0. \tag{21a}$$

Computation of the variable z – Solve for z^{n+1} :

$$\forall K\in\mathcal{M},\quad \frac{1}{\delta t}\,(\rho_K^n z_K^{n+1}-\rho_K^{n-1}z_K^n)+\mathrm{div}_K(\rho^n z^k\boldsymbol{u}^n)=0. \tag{21b}$$

Neutral gas mass fraction computation – Solve for y_N^{n+1} :

$$\forall K\in\mathcal{M},\quad \frac{1}{\delta t}\left[\rho_K^n(y_N)_K^{n+1}-\rho_K^{n-1}(y_N)_K^n\right]+\mathrm{div}_K(\rho^n y_N^k\boldsymbol{u}^n)=0. \tag{21c}$$

Fuel mass fraction computation – Solve for y_F^{n+1} :

$$\forall K\in\mathcal{M},\quad \frac{1}{\delta t}\left[\rho_K^n(y_F)_K^{n+1}-\rho_K^{n-1}(y_F)_K^n\right]+\mathrm{div}_K(\rho^n y_F^k\boldsymbol{u}^n)=-\frac{1}{\varepsilon}\nu_F W_F\,\dot\omega_K^{n+1}. \tag{21d}$$

Product mass fraction computation – Compute y_P^{n+1} given by:

$$\forall K\in\mathcal{M},\quad (y_P)_K^{n+1}=1-(y_F)_K^{n+1}-(y_O)_K^{n+1}-(y_N)_K^{n+1}. \tag{21e}$$

The initial value of the chemical variables is the mean value of the initial condition over the primal cells:

$$\forall K \in \mathcal{M}, \quad G_K^0 = \frac{1}{|K|}\int_K G_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad z_K^0 = \frac{1}{|K|}\int_K z_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$
$$(y_i)_K^0 = \frac{1}{|K|}\int_K (y_i)_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\ i = N, F,$$

where the reduced variable z is the linear combination of y_F and y_O given by Eq. (13).

In Eqs. (21a)–(21d), the discretization of the convection terms is performed by a discrete operator of the form (19). Several choices are possible (and compared in numerical tests) for the evaluation of the value at the face: either an implicit scheme (*i.e.* $k = n+1$) with a first-order upwind space discretization, either an explicit scheme (*i.e.* $k = n$) with a MUSCL or an anti-diffusive space approximation. These latter discretizations are described in Sect. 8 and Sect. 9 of the appendix, respectively.

According to the developments of Sect. 3, the chemical reaction term reads $\dot{\omega}_K^{n+1} = \eta((y_F)_K^{n+1}, z_K^{n+1})\ (G_K^{n+1} - 0.5)^-$ with

$$\eta((y_F)_K^{n+1}, z_K^{n+1}) = \begin{cases} \dfrac{1}{\nu_F W_F}\,(y_F)_K^{n+1} & \text{if } z^{n+1} \le 0, \\ \dfrac{1}{\nu_F W_F}\,(y_F)_K^{n+1} - z_K^{n+1} & \text{otherwise,} \end{cases}$$

and the chemical species mass fractions satisfy the following system, which is equivalent to (21b)–(21e):

$$\frac{1}{\delta t}\left(\rho_K^n (y_i)_K^{n+1} - \rho_K^{n-1}(y_i)_K^n\right) + \mathrm{div}_K(\rho^n y_i^k \boldsymbol{u}^n) = \frac{1}{\varepsilon}\,\zeta_i\,\nu_i\,W_i\,\dot{\omega}^{n+1},$$
$$\text{for } i \in \mathcal{I} \text{ and } K \in \mathcal{M}. \qquad (22)$$

At the continuous level, the last term of equation (21a) may be written

$$\rho_u\,u_f\,|\nabla G| = \boldsymbol{a}\cdot\nabla G = \mathrm{div}(G\,\boldsymbol{a}) - G\,\mathrm{div}(\boldsymbol{a}), \quad \text{with } \boldsymbol{a} = \rho_u\,u_f\,\frac{\nabla G}{|\nabla G|},$$

and we use the same decomposition at the discrete level:

$$|K|\,(\rho_u\,u_f\,|\nabla G|)_K = \sum_{\sigma\in\mathcal{E}(K)} |\sigma|\,(G_\sigma^k - G_K^k)\,\boldsymbol{a}_\sigma^n\cdot\boldsymbol{n}_{K,\sigma},$$

where G_σ^k may be given by one of the three above-mentioned schemes, namely an implicit upwind (with respect to $\boldsymbol{a}^n\cdot\boldsymbol{n}_{K,\sigma}$) scheme, an explicit MUSCL or an explicit anti-diffusive scheme. The flame velocity on σ, $\boldsymbol{a}_\sigma^n$, is evaluated as

$$\boldsymbol{a}_\sigma^n = \rho_u\,u_f\,\frac{(\nabla G)_\sigma^n}{|(\nabla G)_\sigma^n|},$$

where the gradient of G on $\sigma = K|L$ is computed as:

$$(\nabla G)_\sigma = \frac{1}{|K \cup L|}\Big[\sum_{\tau\in\mathcal{E}(K)} |\tau|\, \hat{G}_\tau\, \boldsymbol{n}_{K,\tau} + \sum_{\tau\in\mathcal{E}(L)} |\tau|\, \hat{G}_\tau\, \boldsymbol{n}_{L,\tau}\Big],$$

where $\hat{G}_\tau$ is a second order approximation of G at the center of the face τ.

6 Scheme Conservativity

Let the discrete sensible internal energy be defined by $p_K^n = (\gamma - 1)\,\rho_K^n (e_s)_K^n$ for $K \in \mathcal{M}$ and $0 \le n \le N$. In view of the equation of state (18f), this definition implies $\rho_K^n (h_s)_K^n = \rho_K^n (e_s)_K^n + p_K^n$, for $K \in \mathcal{M}$ and $0 \le n \le N$. The following lemma states that the discrete solutions satisfy a local internal energy balance.

Lemma 1. (Discrete internal energy balance)
A solution to (18)–(21) satisfies the following equality, for any $K \in \mathcal{M}$ and $0 \le n < N$:

$$\frac{1}{\delta t}\big[(\rho e)_K^{n+1} - (\rho e)_K^n\big] + \widetilde{\mathrm{div}}_K(\rho e \boldsymbol{u})^{n+1} + p_K^{n+1}\mathrm{div}_K(\boldsymbol{u})^{n+1} = S_K^{n+1}, \tag{23}$$

where

$$(\rho e)_K^{n+1} = \rho_K^{n+1}(e_s)_K^{n+1} + \rho_K^n \sum_{i\in\mathcal{I}} \Delta h_{f,i}^0 (y_i)_K^{n+1},$$
$$\widetilde{\mathrm{div}}_K(\rho e \boldsymbol{u})^{n+1} = \mathrm{div}_K\Big[(\rho e_s)^{n+1}\boldsymbol{u}^{n+1} + \rho^n\big[\sum_{i\in\mathcal{I}} \Delta h_{f,i}^0\, y_i^k\big]\boldsymbol{u}^n\Big].$$

Proof. We begin by deriving a local sensible internal energy balance, starting from the sensible enthalpy balance (18e) and mimicking the previously given formal passage between these two equations at the continuous level (*i.e.* the passage from Eq. (11) to Eq. (10)). To this purpose, let us write (18e) as $T_1 + T_2 = T_3$ with

$$T_1 = \frac{1}{\delta t}\left[\rho_K^{n+1}\,(h_s)_K^{n+1} - \rho_K^n\,(h_s)_K^n\right] - \frac{1}{\delta t}\,(p_K^{n+1} - p_K^n),$$
$$T_2 = \mathrm{div}_K(\rho h_s \boldsymbol{u})^{n+1} - \big(\boldsymbol{u}\cdot\nabla p\big)_K^{n+1},$$
$$T_3 = (\dot{\omega}_\theta)_K^{n+1} + S_K^{n+1}.$$

Using $\rho_K^\ell (h_s)_K^\ell = \rho_K^\ell (e_s)_K^\ell + p_K^\ell$ for $\ell = n$ and $\ell = n+1$, we easily get

$$T_1 = \frac{1}{\delta t}\left[\rho_K^{n+1}\,(e_s)_K^{n+1} - \rho_K^n\,(e_s)_K^n\right].$$

The term T_2 reads:

$$|K|\, T_2 = \sum_{\sigma \in \mathcal{E}(K)} |\sigma| \left[\rho_\sigma^{n+1} (h_s)_\sigma^{n+1} - p_\sigma^{n+1} + p_K^{n+1}\right] u_{K,\sigma}^{n+1}.$$

If $u_{K,\sigma}^{n+1} > 0$, by definition, $\rho_\sigma^{n+1}(h_s)_\sigma^{n+1} = \rho_K^{n+1}(h_s)_K^{n+1}$ and $p_\sigma^{n+1} = p_K^{n+1}$; otherwise, thanks to the assumptions on the boundary conditions, σ is an internal face and, denoting by L the adjacent cell to K such that $\sigma = K|L$, $\rho_\sigma^{n+1}(h_s)_\sigma^{n+1} = \rho_L^{n+1}(h_s)_L^{n+1}$ and $p_\sigma^{n+1} = p_L^{n+1}$. In both cases, denoting by $(e_s)_\sigma^{n+1}$ the upwind choice for $(e_s)^{n+1}$ at the face σ, we get

$$\rho_\sigma^{n+1}(h_s)_\sigma^{n+1} - p_\sigma^{n+1} = \rho_\sigma^{n+1}(e_s)_\sigma^{n+1},$$

so, finally

$$|K|\, T_2 = \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1}\, (e_s)_\sigma^{n+1} + p_K^{n+1} \sum_{\sigma \in \mathcal{E}(K)} |\sigma|\, u_{K,\sigma}^{n+1}.$$

We thus get the following sensible internal energy balance:

$$\frac{|K|}{\delta t}\left[\rho_K^{n+1}\,(e_s)_K^{n+1} - \rho_K^n\,(e_s)_K^n\right] + \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1}\,(e_s)_\sigma^{n+1} + p_K^{n+1} \sum_{\sigma \in \mathcal{E}(K)} |\sigma|\, u_{K,\sigma}^{n+1} = |K|\left[(\dot{\omega}_\theta)_K^{n+1} + S_K^{n+1}\right], \tag{24}$$

or, using the discrete differential operator formalism,

$$\frac{1}{\delta t}\left[\rho_K^{n+1}\,(e_s)_K^{n+1} - \rho_K^n\,(e_s)_K^n\right] + \mathrm{div}_K(\rho e_s \boldsymbol{u})^{n+1} + p_K^{n+1}\mathrm{div}_K \boldsymbol{u}^{n+1} = (\dot{\omega}_\theta)_K^{n+1} + S_K^{n+1}. \tag{25}$$

We now derive from this relation a discrete (sensible and chemical) internal energy balance. Multiplying the mass fraction balance equations by the corresponding formation enthalpy $(\Delta h_{f,i}^0)_{i \in \mathcal{I}}$ and summing over $i \in \mathcal{I}$ yields:

$$\frac{1}{\delta} t \sum_{i \in \mathcal{I}} \Delta h_{f,i}^0 \left[\rho_K^n (y_i)_K^{n+1} - \rho_K^{n+1}(y_i)_K^n\right] + \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^n \sum_{i \in \mathcal{I}} \Delta h_{f,i}^0\, (y_i)_\sigma^k = \sum_{i \in \mathcal{I}} \Delta h_{f,i}^0\, (\dot{\omega}_i)_K^{n+1} = (\dot{\omega}_\theta)_K^{n+1}.$$

Adding this relation to (24) yields the balance equation (23).

Remark 1. (Positivity of the sensible internal energy) Equation (25) implies that the sensible internal energy remains positive, provided that the right-hand side is non-negative, which is true if $\dot{\omega}_\theta \geq 0$, *i.e.* if the chemical reaction is exothermic. The proof

of this property may be found in [8, Lemma 4.3], and relies on two arguments: first, the convection operator may be recast as a discrete positivity-preserving transport operator thanks to the mass balance, and, second, the pressure p_K^{n+1} vanishes when $(e_s)_K^{n+1}$ vanishes, by the equation of state.

The following local discrete kinetic energy balance holds on the dual mesh (see [8, Lemma 4.1] for a proof).

Lemma 2. (Discrete kinetic energy balance on the dual mesh)
A solution to (18)–(21) satisfies the following equality, for $1 \le i \le d$, $\sigma \in \mathcal{E}_{\mathcal{S}}^{(i)}$ *and* $0 \le n < N$*:*

$$\frac{|D_\sigma|}{\delta t}\big[(e_k)_{\sigma,i}^{n+1} - (e_k)_{\sigma,i}^{n}\big] + \sum_{\varepsilon \in \tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\varepsilon}^n (e_k)_{\varepsilon,i}^{n+1} + |D_\sigma|(\nabla p)_{\sigma,i}^{n+1} u_{\sigma,i}^{n+1} = -R_{\sigma,i}^{n+1}, \quad (26)$$

where

$$(e_k)_{\sigma,i}^{n+1} = \frac{1}{2}\rho_{D_\sigma}^n (u_{\sigma,i}^{n+1})^2 + \frac{\delta t^2}{2\rho_{D_\sigma}^n}\big((\nabla p)_{\sigma,i}^{n+1}\big)^2,$$
$$(e_k)_{\varepsilon,i}^{n+1} = \frac{1}{2}\tilde{u}_{\sigma,i}^{n+1}\tilde{u}_{\sigma',i}^{n+1}, \text{ for } \varepsilon = \sigma|\sigma'$$
$$R_{\sigma,i}^{n+1} = \frac{|D_\sigma|\,\rho_{D_\sigma}^{n-1}}{2\delta t}(\tilde{u}_{\sigma,i}^{n+1} - u_{\sigma,i}^{n})^2.$$

We now derive a kinetic energy balance equation on the primal cells from Relation (26). For the sake of clarity, we make a separate exposition for the Rannacher-Turek case and the MAC case. The case of simplicial discretizations, with the degrees of freedom of the Crouzeix-Raviart element, is an easy extension of the Rannacher-Turek case.

The Rannacher-Turek Case
Since the dual meshes are the same for all the velocity components in this case, we may sum up Eq. (26) over $i = 1, \dots d$ to obtain, for $\sigma \in \mathcal{E}$ and $0 \le n < N$:

$$\frac{|D_\sigma|}{\delta t}\big[(e_k)_{\sigma}^{n+1} - (e_k)_{\sigma}^{n}\big] + \sum_{\varepsilon \in \tilde{\mathcal{E}}(D_\sigma)} F_{\sigma,\varepsilon}^n (e_k)_{\varepsilon}^{n+1} + |D_\sigma|(\nabla p)_{\sigma}^{n+1} \cdot \boldsymbol{u}_{\sigma}^{n+1} = -R_{\sigma}^{n+1}, \quad (27)$$

$$\text{with } (e_k)_{\sigma}^{\ell} = \sum_{i=1}^{d} (e_k)_{\sigma,i}^{\ell}, \text{ for } \ell = n \text{ or } \ell = n+1,$$
$$(e_k)_{\varepsilon}^{n+1} = \sum_{i=1}^{d} (e_k)_{\varepsilon,i}^{n+1} \text{ and } R_{\sigma}^{n+1} = \sum_{i=1}^{d} R_{\sigma,i}^{n+1}.$$

For $K \in \mathcal{M}$, let us define a kinetic energy associated to K and the flux $G_{K,\sigma}^{n+1}$ as follows (see Fig. 2):

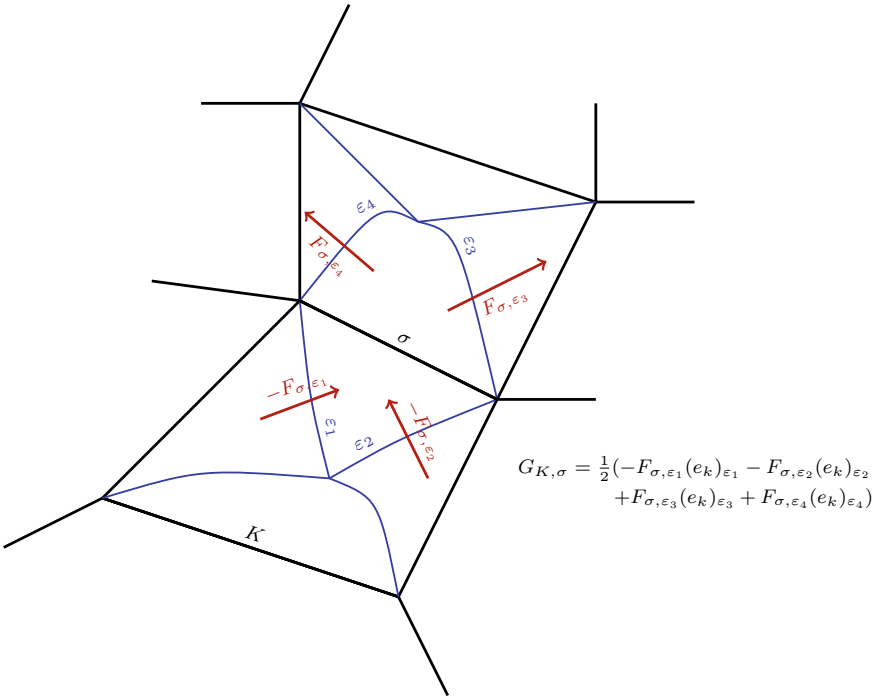


Fig. 2 From fluxes at dual faces to fluxes at primal faces, for the Rannacher-Turek discretization.

$$(e_k)_K^\ell = \frac{1}{2\,|K|} \sum_{\sigma\in\mathcal{E}(K)} |D_\sigma|\,(e_k)_\sigma^\ell,\ \ell = n \text{ or } \ell = n+1,$$
$$G_{K,\sigma}^{n+1} = -\frac{1}{2} \sum_{\varepsilon\in\mathcal{E}(D_\sigma),\varepsilon\subset K} F_{\sigma,\varepsilon}^n\,(e_k)_\varepsilon^{n+1} + \frac{1}{2} \sum_{\varepsilon\in\mathcal{E}(D_\sigma),\varepsilon\not\subset K} F_{\sigma,\varepsilon}^n\,(e_k)_\varepsilon^{n+1}.$$

We easily check that the fluxes $G_{K,\sigma}^{n+1}$ are conservative, in the sense that, for $\sigma = K|L$, $G_{K,\sigma}^{n+1} = -G_{L,\sigma}^{n+1}$. Let us now divide Eq. (27) by 2 and sum over the faces of K. A reordering of the summations, using the conservativity of the mass fluxes through the dual edges and the expression of the discrete pressure gradient, yields:

$$\frac{|K|}{\delta t}\left[(e_k)_K^{n+1} - (e_k)_K^n\right] + \sum_{\sigma\in\mathcal{E}(K)} G_{K,\sigma}^{n+1} + \frac{1}{2} \sum_{\sigma=K|L} |\sigma|\,(p_L^{n+1} - p_K^{n+1})\,u_{K,\sigma}^{n+1} = -R_K^{n+1},$$
$$\text{with } R_K^{n+1} = \frac{1}{2} \sum_{\sigma\in\mathcal{E}(K)} R_\sigma^{n+1}. \tag{28}$$

The MAC Case

Let $1 \le i \le d$, let $K \in \mathcal{M}$, let us denote by σ and σ' the two faces of $\mathcal{E}^{(i)}(K)$, and let us define:

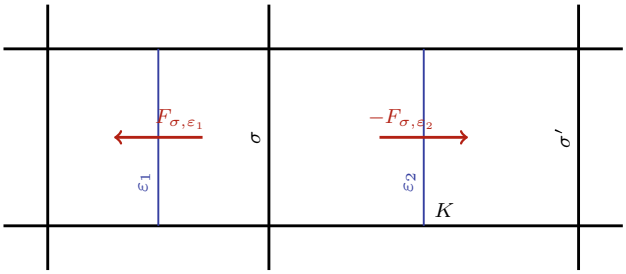


Fig. 3 From fluxes at dual faces to fluxes at primal faces, for the MAC discretization, primal faces parallel to the dual edges, first component of the velocity.

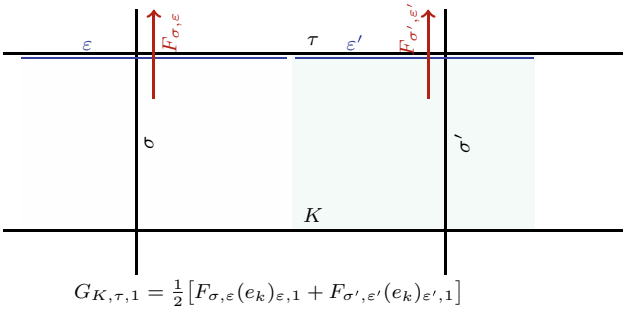


Fig. 4 From fluxes at dual faces to fluxes at primal faces, for the MAC discretization, primal faces orthogonal to the dual edges, first component of the velocity.

$$(e_k)^{\ell}_{K,i} = \frac{1}{2\,|K|}\Big[|D_\sigma|\,(e_k)^{\ell}_{\sigma,i} + |D_{\sigma'}|\,(e_k)^{\ell}_{\sigma',i}\Big], \text{ for } \ell = n \text{ or } \ell = n+1.$$

Case of Primal Faces Parallel to the Dual Faces. Let $\tau = \sigma$ or $\tau = \sigma'$, let ε_1 and ε_2 be the two faces of D_τ perpendicular to $\boldsymbol{e}^{(i)}$, and let ε_2 be the one included in K (see Fig. 3). Then we define

$$G^{n+1}_{K,\tau,i} = \frac{1}{2}\Big[F_{\tau,\varepsilon_1}(e_k)^{n+1}_{\varepsilon_1,i} - F_{\tau,\varepsilon_2}(e_k)^{n+1}_{\varepsilon_2,i}\Big].$$

Case of Primal Faces Orthogonal to the Dual Faces. For $\tau \in \mathcal{E}(K) \setminus \{\sigma, \sigma'\}$, let ε and ε' be such that $\tau \subset (\bar{\varepsilon} \cup \bar{\varepsilon}')$ with ε a face of D_σ and ε' a face of $D_{\sigma'}$ (see Fig. 4).

Then we define

$$G^{n+1}_{K,\tau,i} = \frac{1}{2}\Big[F_{\sigma,\varepsilon}(e_k)^{n+1}_{\varepsilon,i} + F_{\sigma',\varepsilon'}(e_k)^{n+1}_{\varepsilon',i}\Big].$$

Summing Eq. (26) written for σ and for σ' and dividing the result by 2 yields:

$$\frac{|K|}{\delta t}\left[(e_k)_{K,i}^{n+1} - (e_k)_{K,i}^{n}\right] + \sum_{\sigma\in\mathcal{E}(K)} G_{K,\sigma,i}^{n+1} + \frac{1}{2}\sum_{\substack{\sigma\in\mathcal{E}^{(i)}(K)\\ \sigma=K|L}} |\sigma|\,(p_L^{n+1} - p_K^{n+1})\,u_{K,\sigma}^{n+1} = -\frac{1}{2}\left(R_{\sigma,i}^{n+1} + R_{\sigma,i}^{n+1}\right). \tag{29}$$

Now let $(e_k)_K^{\ell} = \sum_{i=1}^{d} (e_k)_{K,i}^{\ell}$, for $\ell = n$ or $\ell = n+1$, and

$$G_{K,\sigma}^{n+1} = \sum_{i=1}^{d} G_{K,\sigma,i}^{n+1}, \text{ for } \sigma \in \mathcal{E}(K).$$

Since only one equation is written for a given face σ of the mesh (for the velocity component i with i such that the normal vector to σ is parallel to $\boldsymbol{e}^{(i)}$), we may define in the MAC case $R_\sigma^{n+1} = R_{\sigma,i}^{n+1}$. Summing Eq. (29) over the space dimension, we finally get

$$\frac{|K|}{\delta t}\left[(e_k)_K^{n+1} - (e_k)_K^{n}\right] + \sum_{\sigma\in\mathcal{E}(K)} G_{K,\sigma}^{n+1} + \frac{1}{2}\sum_{\sigma=K|L} |\sigma|\,(p_L^{n+1} - p_K^{n+1})\,u_{K,\sigma}^{n+1} = -R_K^{n+1}, \text{ with } R_K^{n+1} = \frac{1}{2}\sum_{\sigma\in\mathcal{E}(K)} R_\sigma^{n+1}, \tag{30}$$

which is formally the same equation as Relation (28) (although with a different definition of all the terms in the equation except the pressure gradient).

Remark 2. (On the definition of the cell kinetic energy) Note that, both in the Rannacher-Turek and the MAC case, the cell kinetic energy is not a convex combination of the face kinetic energies, since, on a non-uniform mesh, the equalities $|K| = \frac{1}{2}\sum_{\sigma\in\mathcal{E}(K)} |D_\sigma|$ (Rannacher Turek case) and $|K| = \frac{1}{2}\sum_{\sigma\in\mathcal{E}^{(i)}(K)} |D_\sigma|$ (MAC case) do not hold in general. Consequently, the cell kinetic energy may oscillate from cell to cell while the face kinetic energy does not. Nevertheless, the discrete time derivative of the cell kinetic energy is consistent in the Lax-Wendroff sense, because, despite these oscillations, the cell kinetic energy still converges weakly if the velocity converges.

Equations (28) and (30) suggest a choice for the term S_K^{n+1}, the purpose of which is to compensate the numerical dissipation terms appearing in the kinetic energy balance:

$$S_K^{n+1} = R_K^{n+1}, \text{ for } K \in \mathcal{M} \text{ and } 0 \le n < N. \tag{31}$$

This expression yields a conservative scheme, in the sense that the discrete solutions satisfy a discrete total energy balance without any remainder term (see Eq. (4c) below); as a consequence, the scheme can be proven to be consistent in the Lax-Wendroff sense. However, different definitions are possible (and this latitude may be useful in explicit variants of the scheme, to ensure the positivity of S_K^{n+1}, see Remark 3 below).

We are now in position to state a total energy balance for the scheme.

Theorem 1. (Discrete total energy and stability of the scheme).
A solution to (18)–(21) satisfies the following discrete total energy balance, for any $K \in \mathcal{M}$ and $0 \leq n < N$:

$$\frac{1}{\delta t}\big[(\rho E)_K^{n+1} - (\rho E)_K^n\big] + \widetilde{\mathrm{div}}_K((\rho E + p)\boldsymbol{u})^{n+1} = 0, \tag{32}$$

where

$$(\rho E)_K^\ell = (e_k)_K^\ell + \rho_K^\ell (e_s)_K^\ell + \rho_K^{l-1} \sum_{i\in\mathcal{I}} \Delta h_{f,i}^0 (y_i)_K^\ell, \text{ for } \ell = n \text{ and } \ell = n+1,$$

$$\widetilde{\mathrm{div}}_K((\rho E + p)\,\boldsymbol{u})^{n+1} = \mathrm{div}_K\Big[(\rho e_s)^{n+1}\boldsymbol{u}^{n+1} + \rho^n\big[\sum_{i\in\mathcal{I}} \Delta h_{f,i}^0\, y_i^k\big]\boldsymbol{u}^n\Big]$$
$$+\frac{1}{|K|}\sum_{\sigma\in\mathcal{E}(K)} G_{K,\sigma}^{n+1} + \frac{1}{|K|}\sum_{\sigma=K|L} |\sigma|\,\frac{p_K^{n+1} + p_L^{n+1}}{2}\,u_{K,\sigma}^{n+1}.$$

Let us suppose that e_s^0, ρ^0 and ρ^{-1} are positive. Then, a solution to (18)–(21) satisfies $\rho^n > 0$, $e_s^n > 0$ and the following stability result:

$$E^n = E^0,$$

where, for $0 \leq n \leq N$,

$$E^n = \sum_{K\in\mathcal{M}} |K|(\rho e)_K^n + \frac{1}{2}\sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}_{\mathcal{S}}^{(i)}} |D_\sigma|\,\rho_{D_\sigma}^{n-1}\,(u_{\sigma,i}^n)^2 + \frac{\delta t^2}{2}\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \frac{|D_\sigma|}{\rho_{D_\sigma}^{n-1}} |(\nabla p)_\sigma^n|^2.$$

Proof. The discrete total energy balance equation (32) is obtained by summing the internal energy balance (23) and the kinetic energy balance, *i.e.* Eq. (28) in the Rannacher-Turek case and Eq. (30) for the MAC scheme, and remarking that the numerical dissipation terms in the kinetic energy balance R_K^{n+1} exactly compensate with the corrective terms S_K^{n+1} in the internal energy balance. Then the stability result is obtained by summation over the time steps.

Remark 3. (Consistency of the scheme) The consistency in the Lax-Wendroff sense follows from the conservativity of the scheme (for all balance equations) so, in particular, from the fact that the discrete solutions satisfy the discrete total energy balance (32), thanks to standard (but technical) arguments.

Note however that the consistency of the scheme does not require a strict conservativity, and in particular, variants for the choice (31) of the compensation term in the sensible enthalpy balance are possible; indeed, what is really needed is only that the difference between the dissipation in the kinetic energy balance and its compensation tend to zero in a distributional sense. In practice, this allows a different redistribution of the face residuals to the neighbour primal cells, and this can help to preserve the non-negativity of the compensation term for explicit versions of the scheme.

7 Numerical Tests

At the continuous level, the boundedness of the chemical mass fractions formally implies that, when $\varepsilon \to 0$, the relaxed model converges to the asymptotic one. Indeed, integrating any of the reactive species mass balance equations with respect to time and space, we observe that $||\dot{\omega}||_{L^1(\Omega\times(0,T))}$ tends to zero as ε, and thus two separate zones appear: a zone characterized by $G < 0.5$ where the reaction is complete, and a zone corresponding to $G \geq 0.5$, where no reaction has occured.

A closed form of the solution of the Riemann problem for the asymptotic model is available [1]. In order to perform numerical tests, a Riemann problem with initial conditions such that the analytic solution has the profile presented in Fig. 5 is chosen.

Moreover, the selected configuration imposes zero amplitude for the contact discontinuity and the left non linear wave, thus the solution consists of three different constant states: $\boldsymbol{W}_R^*$, $\boldsymbol{W}^{**}$ and $\boldsymbol{W}_R$. The right state corresponds to a stoichiometric mixture of hydrogen and air (so the molar fractions of Hydrogen, Oxygen and Nitrogen are 2/7, 1/7 and 4/7 respectively) at rest, at the pressure $p = 9.9\,10^4$ Pa and the temperature $T = 283°$ K. The velocity is supposed to be zero in the left state, which is sufficient to determine the solution. Physically, speaking, supposing that the initial discontinuity lies at $x = 0$, this situation corresponds to the left part of a (symmetrical) constant velocity plane deflagration starting at $x = 0$. The flame velocity is $u_f = 63$ m/s and the formation enthalpies are zero except for the product

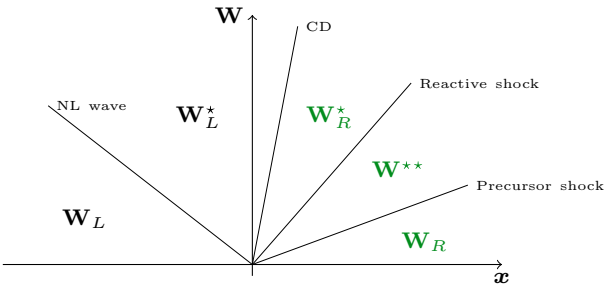


Fig. 5 The analytic solution of the numerical test configuration.

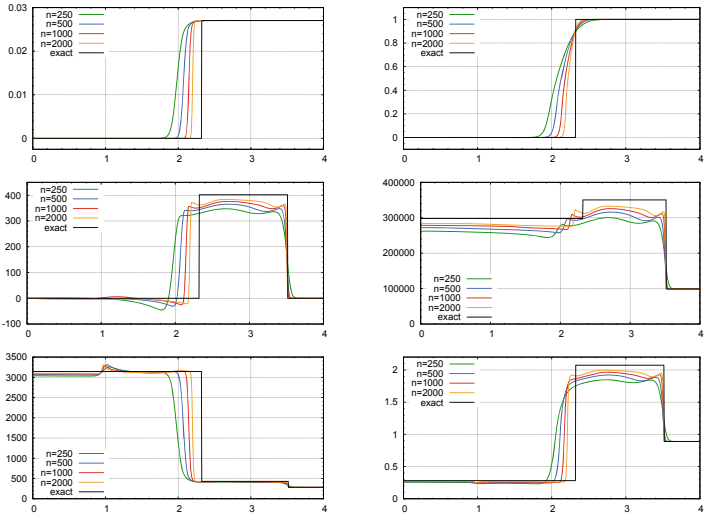


Fig. 6 Upwind scheme – From top left to bottom right, fuel mass fraction, G, velocity, pressure, temperature and density at $t = 0.005$, as a function of the space variable.

(*i.e.* steam), with $\Delta h^0_{f,O} = -13.255\,10^6$ J (Kg K)$^{-1}$. The quantity ρ_u is the analytical density in the intermediate state (so the total velocity of the flame brush is equal to the sum of u_f and the material velocity on the right side of the reactive shock, see [1]). The computation is initialized by the analytical solution at $t = 0.002$ and the final time is $t = 0.005$. The computational domain is the interval $(0, 4.5)$.

The numerical tests performed aim at checking the convergence of the scheme to such a solution, which in fact may result from two different properties: the convergence of the relaxed model to the asymptotic model when ε tends to zero, and the convergence of the scheme towards a numerical solution when the time and space steps tend to zero. To this purpose, we choose ε proportional to the space step and make it tend to zero, with a constant CFL number. We test the scheme behaviour with three different discretizations of the convection operator in the chemical mass species balances: the standard upwind scheme, a MUSCL-like discretization which is an extension to variable density flows of the scheme proposed in [15] and is described in Appendix 8, and a first-order anti-diffusive scheme which is an adaptation to our setting of the scheme proposed in [5]; we detail it in Appendix 9 for the sake of completeness.

Results obtained at $t = 0.005$ with the upwind scheme, the MUSCL-like scheme and the anti-diffusive scheme, for increasingly refined meshes, are shown on Fig. 6, Fig. 7 and Fig. 8 respectively, together with the analytical solution. The expected convergence is indeed observed but, with the upwind discretization, the rate of convergence is poor. This seems to be due to the interaction between the numerical diffusion of the upwind scheme, which artificially introduces unburnt reactive chemical species into the burnt zone, and the stiffness of the reaction term. As expected

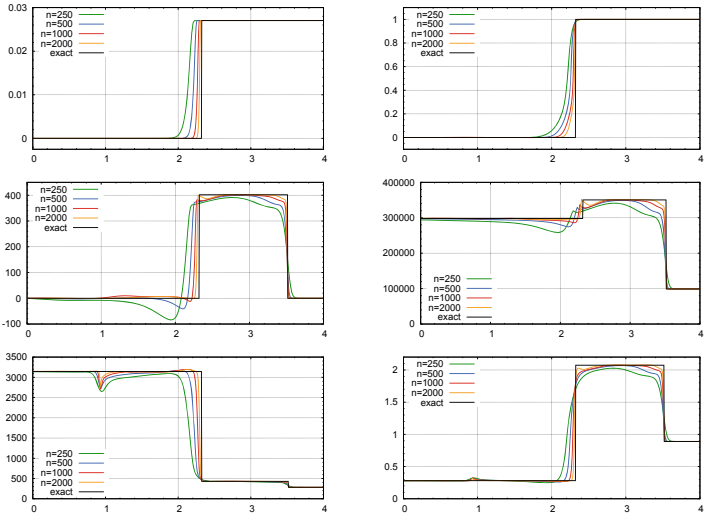


Fig. 7 MUSCL scheme – From top left to bottom right, fuel mass fraction, G, velocity, pressure, temperature and density at $t = 0.005$, as a function of the space variable.

Table 1 L^1 norm of the error between the discrete and continuous solutions for the various schemes - Black: upwind scheme, blue: MUSCL scheme, orange: anti-diffusive scheme; $h_0 = 4.5/250$ is the size of the least refined mesh.

h	$\|\|p - p_{ex}\|\|_{L^1} \times 10^{-4}$			$\|\|\boldsymbol{u} - \boldsymbol{u}_{ex}\|\|_{L^1} \times 10^{-2}$			$\|\|\rho - \rho_{ex}\|\|_{L^1} \times 10$		
h_0	16.5	7.26	4.59	2.17	1.56	1.07	7.69	3.71	2.74
$\frac{h_0}{2}$	12.5	3.88	2.43	1.64	0.787	0.579	6.16	2.23	1.65
$\frac{h_0}{4}$	9.66	2.05	1.38	1.23	0.471	0.371	4.73	1.26	0.913
$\frac{h_0}{8}$	7.58	1.17	0.708	0.958	0.263	0.175	3.63	0.691	0.476
$\frac{h_0}{20}$	5.78	0.673	0.375	0.728	0.160	0.103	2.77	0.382	0.267
$\frac{h_0}{40}$	4.31	0.414	0.194	0.543	0.0786	0.0458	2.03	0.201	0.134

in such a case, the results are significantly improved by the use of a less diffusive scheme for the chemical species balance equations. Indeed, passing from the upwind to the MUSCL-like and to the anti-diffusive discretization improves the accuracy of the scheme, as may be observed in Fig. 9, where the results obtained by the three discretizations for a regular mesh composed of 500 cells are plotted together with the continuous solution.

This observation is comforted by the measure, in L^1-norm, of the difference between the discrete and continuous solutions, see Table 1. For every mesh and variable, the anti-diffusive scheme is the most accurate and the upwind one the least. The calculated order of convergence is close to 0.5 for the upwind scheme, and to 1 for the MUSCL-like and anti-diffusive schemes.

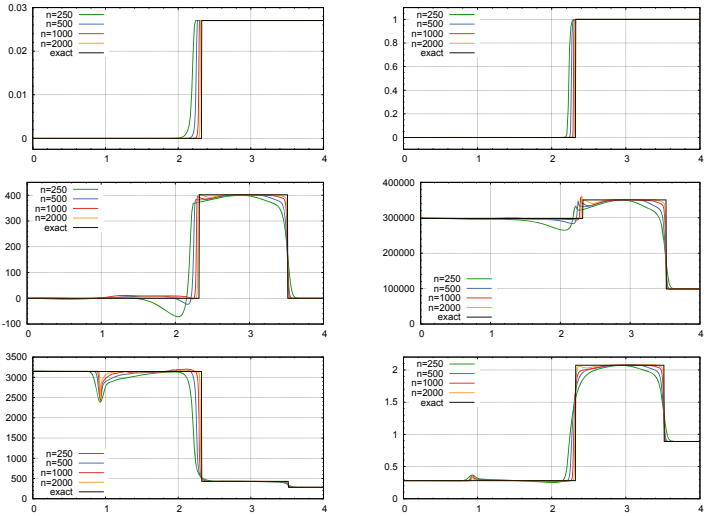


Fig. 8 Anti-diffusive scheme – From top left to bottom right, fuel mass fraction, G, velocity, pressure, temperature and density at $t = 0.005$, as a function of the space variable.

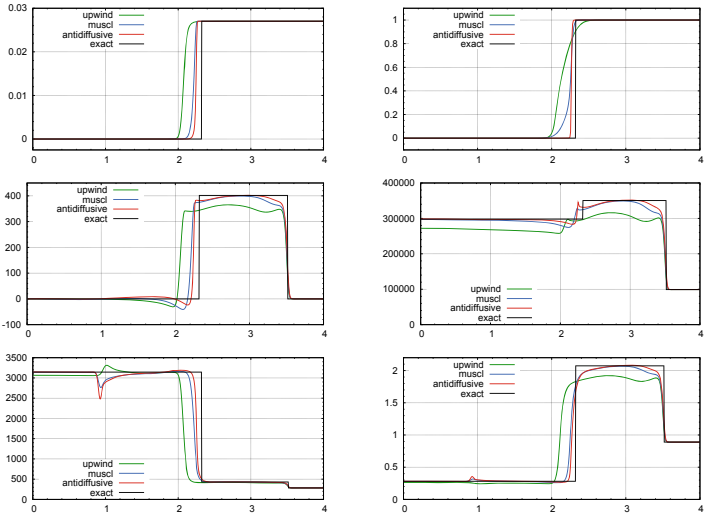


Fig. 9 Comparison of the solutions obtained with the upwind, MUSCL and anti-diffusive scheme – From top to bottom, fuel mass fraction, G, velocity, pressure, temperature and density at $t = 0.005$, as a function of the space variable. Results obtained with a regular mesh composed of $n = 500$ cells.

8 Appendix A: The MUSCL Scheme

The MUSCL discretization of the convection operators of the chemical species balance and G-equation closely follows the technique proposed in [15]. To present this discretization, we consider the following system of equations:

$$\begin{aligned}&\partial_t \rho + \operatorname{div}(\rho \boldsymbol{u}) = 0,\\ &\partial_t (\rho \boldsymbol{u}) + \operatorname{div}(\rho \boldsymbol{u} y) = 0.\end{aligned}$$

We suppose for short that this system is complemented by impermeability boundary conditions, *i.e.* that the normal velocity, both at the continuous and the discrete level, vanishes on the boundary of the computational domain.

The discretization of the above system reads:

$$\begin{aligned}\forall K \in \mathcal{M},\ &\frac{\rho_K^{n+1} - \rho_K^n}{\delta t} + \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1} = 0,\\ &\frac{\rho_K^{n+1} y_K^{n+1} - \rho_K^n y_K^n}{\delta t} + \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1} y_\sigma^n = 0.\end{aligned}$$

For any $\sigma \in \mathcal{E}$, the procedure consists in three steps:

- calculate a tentative value for y_σ as a linear interpolate of nearby values,
- calculate an interval for y_σ which guarantees the stability of the scheme,
- project the tentative value y_σ on this stability interval.

For the tentative value of y_σ, let us choose some real coefficients $(\alpha_K^\sigma)_{K \in \mathcal{M}}$ such that

$$\boldsymbol{x}_\sigma = \sum_{K \in \mathcal{M}} \alpha_K^\sigma \boldsymbol{x}_K, \qquad \sum_{K \in \mathcal{M}} \alpha_K^\sigma = 1.$$

The coefficients used in this interpolation are chosen in such a way that as few as possible cells, to be picked up in the closest cells to σ, take part. For example, for $\sigma = K|L$ and if $\boldsymbol{x}_K$, $\boldsymbol{x}_\sigma$, $\boldsymbol{x}_L$ are aligned, only two non-zero coefficients exist in the family $(\alpha_K^\sigma)_{K \in \mathcal{M}}$, namely α_K^σ and α_L^σ. Then, these coefficients are used to calculate the tentative value of y_σ by

$$y_\sigma = \sum_{K \in \mathcal{M}} \alpha_K^\sigma y_K.$$

The construction of the stability interval must be such that the following property holds:

$$\begin{aligned}&\forall K \in \mathcal{M},\ \forall \sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\text{int}},\ \exists \beta_K^\sigma \in [0,1] \text{ and } M_K^\sigma \in \mathcal{M} \text{ such that}\\ &y_\sigma^n - y_K^n = \begin{cases} \beta_K^\sigma\,(y_K^n - y_{M_K^\sigma}^n), & \text{if } F_{K,\sigma}^{n+1} \geq 0,\\ \beta_K^\sigma\,(y_{M_K^\sigma}^n - y_K^n), & \text{otherwise.}\end{cases}\end{aligned} \tag{33}$$

Indeed, under this latter hypothesis and a CFL condition, the scheme preserves the initial bounds of y.

Remark 4. Note that, in Assumption (33), only internal faces are considered, since the fluxes through external faces are supposed to vanish. However, the present discussion may easily be generalized to cope with convection fluxes entering the domain.

Definition 1. The so-called CFL number reads for any $0 \leq n \leq N$:

$$\mathrm{CFL}^{n+1} = \max_{K \in \mathcal{M}} \left\{ \frac{\delta t}{\rho_K^{n+1} |K|} \sum_{\sigma \in \mathcal{E}(K)} \left| F_{K,\sigma}^{n+1} \right| \right\}.$$

Lemma 3. *Let us suppose that* $\mathrm{CFL}^{n+1} \leq 1$. *For* $K \in \mathcal{M}$, *let us note by* $\mathcal{V}(K)$ *the union of the set of cells* M_K^σ, $\sigma \in \mathcal{E}(K) \cap \mathcal{E}_{\mathrm{int}}$ *such that* (33) *holds. Then* $\forall K \in \mathcal{M}$, *the value of* y_K^{n+1} *is a convex combination of* $\{y_K^n, (y_M^n)_{M \in \mathcal{V}(K)}\}$.

Proof. The discrete mass balance equation yields:

$$\rho_K^n = \rho_K^{n+1} + \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1}.$$

Replacing this expression of ρ_K^n in the discrete balance equation of y and using the relations provided by (33), we obtain:

$$\begin{aligned}
\rho_K^{n+1} y_K^{n+1} &= \rho_K^n y_K^n - \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1} y_\sigma^n \\
&= \rho_K^{n+1} y_K^n - \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma}^{n+1} (y_\sigma^n - y_K^n) \\
&= \rho_K^{n+1} y_K^n - \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} \left(F_{K,\sigma}^{n+1}\right)^+ (y_\sigma^n - y_K^n) + \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} \left(F_{K,\sigma}^{n+1}\right)^- (y_\sigma^n - y_K^n) \\
&= \rho_K^{n+1} y_K^n - \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} \left(F_{K,\sigma}^{n+1}\right)^+ \beta_K^\sigma (y_K^n - y_{M_K^\sigma}^n) + \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} \left(F_{K,\sigma}^{n+1}\right)^- \beta_K^\sigma (y_{M_K^\sigma}^n - y_K^n).
\end{aligned}$$

This relation yields

$$y_K^{n+1} = y_K^n \left(1 - \frac{\delta t}{\rho_K^{n+1} |K|} \sum_{\sigma \in \mathcal{E}(K)} \beta_K^\sigma \left| F_{K,\sigma}^{n+1} \right| \right) + \frac{\delta t}{|K|} \sum_{\sigma \in \mathcal{E}(K)} y_{M_K^\sigma}^n \beta_K^\sigma \left| F_{K,\sigma}^{n+1} \right|,$$

which concludes the proof under the hypothesis that $\mathrm{CFL} \leq 1$.

We now need to reformulate (33) in order to construct the stability interval. Let $\sigma \in \mathcal{E}$, let us denote by V^- and V^+ the upstream and downstream cell separated by σ, and by $\mathcal{V}_\sigma(V^-)$ and $\mathcal{V}_\sigma(V^+)$ two sets of neighbouring cells of V^- and V^+ respectively, and let us suppose:

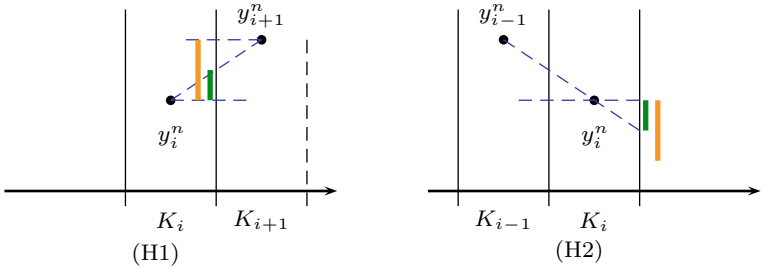


Fig. 10 Conditions (H1) and (H2) in 1D.

$$\text{(H1)} - \exists M \in \mathcal{V}_\sigma(V^+) \text{ s.t. } u^n_\sigma \in |[u^n_M, u^n_M + \frac{\zeta^+}{2}(u^n_{V^+} - u^n_M)]|,$$
$$\text{(H2)} - \exists M \in \mathcal{V}_\sigma(V^-) \text{ s.t. } u^n_\sigma \in |[u^n_{V^-}, u^n_{V^-} + \frac{\zeta^-}{2}(u^n_{V^-} - u^n_M)]|,$$

where for $a,\ b \in \mathbb{R}$, we denote by $|[a, b]|$ the interval $\{\alpha a + (1-\alpha)b,\ \alpha \in [0, 1]\}$, and ζ^+ and ζ^- are two numerical parameters lying in the interval $[0, 2]$.

Conditions (H1)-(H2) and (33) are linked in the following way: let $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}(K)$. If $F^n_{K,\sigma} \leq 0$, *i.e.* K is the downstream cell for σ, denoted above by V^+, since $\zeta^+ \in [0, 2]$, condition (H1) yields that there exists $M \in \mathcal{M}$ such that $u^n_\sigma \in |[u^n_K, u^n_M]|$, which is (33). Otherwise, *i.e.* if $F^n_{K,\sigma} \geq 0$ and K is the upstream cell for σ, denoted above by V^-, condition (H2) yields that there exists $M \in \mathcal{M}$ such that $y^n_\sigma \in |[y^n_K, 2y^n_K - y^n_M]|$, so $y^n_\sigma - y^n_K \in |[0, y^n_K - y^n_M]|$, which is once again (33).

Remark 5. For $\sigma \in \mathcal{E}$, if $V^- \in \mathcal{V}_\sigma(V^+)$, the upstream choice $y^n_\sigma = y^n_{V^-}$ always satisfies the conditions (H1)-(H2), and is the only one to satisfy them if we choose $\zeta^- = \zeta^+ = 0$.

Remark 6. (1D case) Let us take the example of an interface σ separating K_i and K_{i+1} in a 1D case (see Fig. 10 for the notations), with a uniform meshing and a positive advection velocity, so that $V^- = K_i$ and $V^+ = K_{i+1}$. In 1D, a natural choice is $\mathcal{V}_\sigma(K_i) = \{K_{i-1}\}$ and $\mathcal{V}_\sigma(K_{i+1}) = \{K_i\}$. On Fig. 10, we sketch: on the left, the admissible interval given by (H1) with $\zeta^+ = 1$ (green) and $\zeta^+ = 2$ (orange); on the right, the admissible interval given by (H2) with $\zeta^- = 1$ (green) and $\zeta^- = 2$ (orange). The parameters ζ^- and ζ^+ may be seen as limiting the admissible slope between $(\boldsymbol{x}_i, y^n_i)$ and $(\boldsymbol{x}_\sigma, y^n_\sigma)$ (with $\boldsymbol{x}_i$ the abscissa of the mass centre of K_i and $\boldsymbol{x}_\sigma$ the abscissa of σ), with respect to a left and right slope, respectively. For $\zeta^- = \zeta^+ = 1$, one recognizes the usual minmod limiter (*e.g.* [7, Chapter III]). Note that, since, on the example depicted on Fig. 10, the discrete function y^n has an extremum in K_i, the combination of the conditions (H1) and (H2) imposes that, as usual, the only admissible value for y^n_σ is the upwind one.

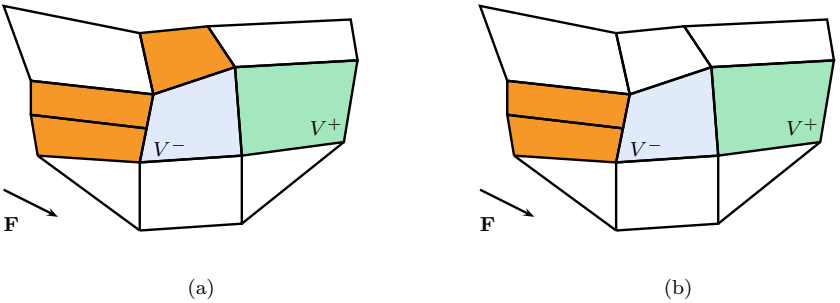


Fig. 11 Notations for the definition of the limitation process. In orange, control volumes of the set $\mathcal{V}_\sigma(V^-)$ for $\sigma = V^-|V^+$, with a constant advection field $\mathbf{F}$: upwind cells (a) or opposite cells (b).

Finally, we need to specify the choice of the sets $\mathcal{V}_\sigma(V^-)$ and $\mathcal{V}_\sigma(V^+)$. Here, we just set $\mathcal{V}_\sigma(V^+) = \{V^-\}$; such a choice guarantees that at least the upstream choice is in the intersection of the intervals defined by (H1) and (H2), as explained in Remark 6. The set $\mathcal{V}_\sigma(V^-)$ may be defined in two different ways (*cf.* Fig. 11):

– as the "upstream cells" to V^-, *i.e.*

$$\mathcal{V}_\sigma(V^-) = \{L \in \mathcal{M},\ L \text{ shares a face } \sigma \text{ with } V^- \text{ and } F_{V^-,\sigma} \leq 0\},$$

– when this makes sense (*i.e.* with a mesh obtained by Q_1 mappings from the $(0,1)^d$ reference element), the opposite cells to σ in V^- are chosen. Note that for a structured mesh, this choice allows to recover the usual minmod limiter.

9 Appendix B: An Anti-diffusive Scheme

The scheme proposed in [5] by of B. Després and F. Lagoutière for the constant velocity advection problem presents some interesting properties in one space dimension (and may be extended to structured multi-dimensional meshes using alternate directions techniques); in particular, it notably limits the numerical diffusion. We extend here this scheme to work with unstructured meshes for which the "opposite cell to a face" (in the sense introduced in the previous section) may be defined and with a variable density. With the same notations as in the previous section, for $\sigma \in \mathcal{E}_{\text{int}}$, $\sigma = K|L$ with $F_{K,\sigma}^{n+1} \geq 0$,

– the tentative value for y_σ is chosen as the downwind value, *i.e.* $y_\sigma^n = y_L^n$,
– Then we project y_σ^n on the interval

$$I_\sigma = \left[y_K^n,\ y_K^n + \frac{1-\nu}{\nu}(y_K - y_M)\right], \quad \nu = \frac{|F_{K,\sigma}^{n+1}|\,\delta t}{\rho_K^{n+1}\,|K|},$$

where $M \in \mathcal{M}$ is the control volume which stands at the opposite side of K with respect to L.

The original scheme presented in [5] is recovered by this formulation for the one-dimensional constant velocity convection equation. Note however, that if the space dimension is greater than one, the above limitation may be not sufficient to reserve the maximum principle.

References

1. Beccantini, A., Studer, E.: The reactive Riemann problem for thermally perfect gases at all combustion regimes. Int. J. Numer. Meth. Fluids **64**, 269–313 (2010)
2. CALIF^{3}S: A software components library for the computation of reactive turbulent flows. https://gforge.irsn.fr/gf/project/isis
3. Ciarlet, P.G.: Basic error estimates for elliptic problems. In: Ciarlet, P., Lions, J. (eds.) Handbook of Numerical Analysis, Vol. II, pp. 17–351. North Holland (1991)
4. Crouzeix, M., Raviart, P.: Conforming and nonconforming finite element methods for solving the stationary Stokes equations. RAIRO Série Rouge **7**, 33–75 (1973)
5. Després, B., Lagoutière, F.: Contact discontinuity capturing scheme for linear advection and compressible gas dynamics. J. Sci. Comput. **16**, 479–524 (2002)
6. Gastaldo, L., Herbin, R., Kheriji, W., Lapuerta, C., Latché, J.C.: Staggered discretizations, pressure correction schemes and all speed barotropic flows. In: Finite Volumes for Complex Applications VI - Problems & Perspectives - Prague, Czech Republic, vol. 2, pp. 39–56. Springer (2011)
7. Godlewski, E., Raviart, P.A.: Numerical Approximation of Hyperbolic Systems of Conservation Laws. No. 118 in Applied Mathematical Sciences. Springer, New York (1996)
8. Grapsas, D., Herbin, R., Kheriji, W., Latché, J.C.: An unconditionally stable finite element-finite volume pressure correction scheme for the compressible Navier-Stokes equations. SMAI J. Comput. Math. **2**, 51–97 (2016)
9. Harlow, F., Amsden, A.: A numerical fluid dynamics calculation method for all flow speeds. J. Comput. Phys. **8**, 197–213 (1971)
10. Harlow, F., Welsh, J.: Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. Phys. Fluids **8**, 2182–2189 (1965)
11. Herbin, R., Kheriji, W., Latché, J.C.: On some implicit and semi-implicit staggered schemes for the shallow water and Euler equations. Math. Modelling Numer. Anal. **48**, 1807–1857 (2014)
12. Herbin, R., Latché, J.C.: Kinetic energy control in the MAC discretisationof the compressible Navier-Stokes equations. Int. J. Finite Vol. **7** (2010)
13. Larrouturou, B.: How to preserve the mass fractions positivity when computing compressible multi-component flows. J. Comput. Phys. **95**, 59–84 (1991)
14. Peters, N.: Turbulent Combustion. Cambridge University Press, Cambridge Monographs of Mechanics (2000)
15. Piar, L., Babik, F., Herbin, R., Latché, J.C.: A formally second order cell centered scheme for convection-diffusion equations on general grids. Int. J. Numer. Meth. Fluids **71**, 873–890 (2013)
16. Poinsot, T., Veynante, D.: Theoretical and Numerical Combustion. Editions R.T Edwards Inc. (2005)
17. Rannacher, R., Turek, S.: Simple nonconforming quadrilateral Stokes element. Numer. Methods Partial Differ. Equ. **8**, 97–111 (1992)
18. Zimont, V.: Gas premixed combustion at high turbulence. Turbulent flame closure combustion model. Experimental Thermal Fluid Sci. **21**, 179–186 (2000)

New Invariant Domain Preserving Finite Volume Schemes for Compressible Flows

Mária Lukáčová-Medviďová, Hana Mizerová, and Bangwei She

Abstract We present new invariant domain preserving finite volume schemes for the compressible Euler and Navier–Stokes–Fourier systems. The schemes are entropy stable and preserve positivity of density and internal energy. More importantly, their convergence towards a strong solution of the limit system has been proved rigorously in [9, 11]. We will demonstrate their accuracy and robustness on a series of numerical experiments.

Keywords Compressible Euler and Navier–Stokes–Fourier systems · Finite volume methods · Invariant domain preserving properties · Entropy stability · Convergence

1 Introduction

Numerical simulations of compressible flows find their applications in many everyday problems, ranging from engineering, oceanography, meteorology to

M. Lukáčová-Medviďová
Institute of Mathematics, Johannes Gutenberg-University Mainz,
Staudingerweg 9, 55 128 Mainz, Germany
e-mail: lukacova@uni-mainz.de

H. Mizerová · B. She
Institute of Mathematics of the Czech Academy of Sciences,
Žitná 25, 115 67 Praha 1, Czech Republic
e-mail: she@math.cas.cz

H. Mizerová (✉)
Department of Mathematical Analysis and Numerical Mathematics, Comenius University,
Mlynská dolina, 842 48 Bratislava, Slovakia
e-mail: hana.mizerova@fmph.uniba.sk

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*,
SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2_6

hemodynamics. Over the years a large variety of powerful numerical schemes has been developed. Let us point out a few well-established and practical schemes, e.g., [1, 5, 6, 12, 16, 18, 19, 23, 25]. Despite of their practical success the rigorous numerical analysis, in particular, in multiple space dimensions, is still open in general.

In [13, 14] the concept of *invariant domain preserving schemes* for hyperbolic conservation laws has been introduced. These methods satisfy some important structure preserving properties, such as positivity of some quantities, entropy production or the minimum entropy principle. In our recent works [9–11] we have proposed new finite volume schemes for the compressible Euler equations of gas dynamics, compressible Navier–Stokes and Navier–Stokes–Fourier equations, respectively. Our new finite volume methods belong to the class of the invariant domain preserving schemes. Their properties further allowed us to study the convergence of the schemes rigorously. More precisely, we proved a nonlinear variant of the *Lax equivalence theorem*: a consistent numerical scheme is convergent if and only if it is stable.

Of course, the compressible Euler and Navier–Stokes–Fourier equations are truly nonlinear, thus we have to overcome difficulties arising due to nonlinear terms. To this goal, we apply a concept of dissipative measure–valued solutions developed in [2, 3, 8] for the above mentioned systems, respectively. Indeed, the Young measures which are the space-time parametrized probability measures replace the linearity setting. They allow us to pass to the limit in nonlinear terms and show the convergence of our finite volume schemes. A limit is in general only a measure, more precisely a dissipative measure–valued solution. We refer a reader to [2, 3, 8] and [9–11] for more details on its definition.

A crucial ingredient of our convergence analysis is the fact that we have the weak-strong uniqueness principle for all systems mentioned above. More precisely, if the strong solution exists our dissipative measure–valued solution coincides with the former on its lifespan. Consequently, we get the strong convergence of our numerical solutions to the strong solution in appropriate Lebesgue spaces. The main aim of this paper is to illustrate experimentally the behavior of our new invariant domain preserving finite volume schemes for compressible fluids, namely for the Euler and the Navier–Stokes–Fourier systems, cf. [9, 11].

The gas dynamics of inviscid compressible flows is governed by the Euler equations

$$\begin{aligned} \partial_t \varrho + \mathrm{div}_x \boldsymbol{m} &= 0, \\ \partial_t \boldsymbol{m} + \mathrm{div}_x (\boldsymbol{m} \otimes \boldsymbol{u}) + \nabla_x p &= 0, \\ \partial_t E + \mathrm{div}_x ((E + p)\boldsymbol{u}) &= 0, \end{aligned} \tag{1}$$

where ϱ, p, $\boldsymbol{u}$, $\boldsymbol{m} = \varrho \boldsymbol{u}$, and E represent the density, pressure, velocity, momentum and the total energy of a fluid, respectively. Taking into account the viscous and heat conducting effects yields the Navier–Stokes–Fourier equations

$$\partial_t \varrho + \mathrm{div}_x \boldsymbol{m} = 0,$$

$$\partial_t \boldsymbol{m} + \mathrm{div}_x(\boldsymbol{m} \otimes \boldsymbol{u}) + \nabla_x p = \mathrm{div}_x \mathbb{S}(\mathbf{D}(\boldsymbol{u})),$$
$$\partial_t(\varrho e) + \mathrm{div}_x(\varrho e \boldsymbol{u}) - \mathrm{div}_x(\kappa \nabla_x \vartheta) = \mathbb{S}(\mathbf{D}(\boldsymbol{u})) : \nabla_x \boldsymbol{u} - p\mathrm{div}_x \boldsymbol{u}, \tag{2}$$

where the viscous stress tensor $\mathbb{S}$ reads

$$\mathbb{S}(\mathbf{D}(\boldsymbol{u})) = 2\mu \mathbf{D}(\boldsymbol{u}) + \lambda \mathrm{div}_x \boldsymbol{u} \mathbb{I}, \quad \mathbf{D}(\boldsymbol{u}) = \frac{\nabla_x \boldsymbol{u} + \nabla_x \boldsymbol{u}^T}{2}.$$

The systems Eq. 1 and Eq. 2 are closed by the standard pressure law for a perfect gas $p = p(\varrho, \vartheta) = \varrho\vartheta$, ϑ is the temperature. Denoting further e the specific internal energy, s the physical entropy, $\gamma > 1$ the adiabatic coefficient and $c_v = \frac{1}{\gamma - 1}$ the specific heat at constant volume we have

$$e(\varrho, \vartheta) = c_v \vartheta, \ s(\varrho, \vartheta) = \log\left(\frac{\vartheta^{c_v}}{\varrho}\right) = \frac{1}{\gamma - 1} \log\left(\frac{p}{\varrho^\gamma}\right).$$

The total energy $E = \frac{1}{2}\frac{\boldsymbol{m}^2}{\varrho} + \varrho e$ consists of the kinetic energy and the internal energy.

Both systems Eq. 1 and Eq. 2 are solved in the time-space cylinder $(0, T) \times \boldsymbol{\Omega}$, $\boldsymbol{\Omega} \subset R^d$, $d = 2, 3$. We assume that these systems are accompanied with appropriate boundary conditions: either the periodic boundary conditions when the domain $\boldsymbol{\Omega}$ is identified with a flat torus, or the no-flux boundary conditions,

$$\boldsymbol{u}|_{\partial\Omega} \cdot \boldsymbol{n} = 0, \ \nabla_x \vartheta \cdot \boldsymbol{n} = 0$$

in the case of the Euler equations, see Eq. 1, or the no–slip boundary conditions,

$$\boldsymbol{u}|_{\partial\Omega} = 0$$

for the Navier–Stokes–Fourier system, see Eq. 2. To close the formulation of the problem we impose the initial conditions

$$\mathbf{U}(0) = \mathbf{U}_0, \ \text{with } \varrho_0 > 0 \text{ and } E_0 - \frac{1}{2}\frac{|\boldsymbol{m}_0|^2}{\varrho_0} > 0, \tag{3}$$

where $\mathbf{U} = (\varrho, \boldsymbol{m}, E)$ or $\mathbf{U} = (\varrho, \boldsymbol{m}, \vartheta)$ for the Euler and the Navier–Stokes–Fourier equations, respectively.

2 Finite Volume Schemes

We start by introducing the mesh, space discretization and suitable discrete spaces.

2.1 Mesh and Space Discretization

Primary Grid. We suppose the physical space to be a polyhedral domain $\Omega \subset R^d$, $d = 2, 3$, that is decomposed into compact elements,

$$\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K.$$

The elements K are sharing either a common face, edge, or vortex.

They can be chosen to be triangular, rectangular, or any combination of them. The primary mesh $\mathcal{T}_h$ is assumed to satisfy the standard regularity assumptions, cf. [4, 7]. The set of all faces is denoted by Σ, while the set of faces on the boundary is denoted by Σ_{ext}, and the set of interior faces by $\Sigma_{int} = \Sigma \setminus \Sigma_{ext}$. Note that there is no boundary if the flow is periodic:

$$\Sigma_{ext} = \emptyset \text{ and } \Sigma_{int} = \Sigma.$$

Each face is associated with an outer normal vector $\boldsymbol{n}$. Let $|K|$, $|\sigma|$ be the Lebesgue measure of an element K and a face σ, respectively. We shall suppose

$$|K| \approx h^d, \ |\sigma| \approx h^{d-1} \text{ for any } K \in \mathcal{T}_h, \ \sigma \in \Sigma.$$

The parameter $h \in (0, 1)$ is the maximum element size, i.e., the size of the mesh $\mathcal{T}_h$.

For the discretization of the Navier–Stokes–Fourier system, see Eq. 2, we additionally require the primary grid $\mathcal{T}_h$ to satisfy the following property: there is a family of control points $P_h = \{x_K \mid x_K \in K, K \in \mathcal{T}_h\}$, such that the segment $[x_K, x_L]$ for any adjacent elements K and L is perpendicular to their common face $\sigma = K \cap L$. We denote by $d_\sigma = (x_K, x_L)$ the Euclidean distance between the points x_K and x_L in R^d. This requirement is naturally satisfied by any rectangular mesh with P_h being the set of gravity centers of all elements. For a triangular mesh, we can use the well-centered mesh [24], where P_h is the set of circumcenters of all elements.

Dual Grid. For the theoretical analysis of our finite volume scheme for the Navier–Stokes–Fourier system it is convenient to introduce a dual mesh $\mathcal{D}_h$. A dual cell D_σ associated to a face $\sigma = K \cap L$ is defined as $D_\sigma = D_{\sigma,K} \cup D_{\sigma,L}$, where $D_{\sigma,K}$ ($D_{\sigma,L}$) is a triangle (tetrahedron) built by x_K and the common vertices of K and L, see Fig. 1 for a two-dimensional example.

Discrete Function Spaces. We denote by Q_h and W_h the set of piecewise constant functions on the primary grid $\mathcal{T}_h$ and the dual grid $\mathcal{D}_h$, respectively. Moreover, $\boldsymbol{v}_h \in Q_h$ (resp. $\boldsymbol{v}_h \in W_h$) means that each component of $\boldsymbol{v}_h$ belongs to Q_h (resp. W_h). Further, for a piecewise continuous function v, whenever $x \in \sigma \in \Sigma_{int}$, we define

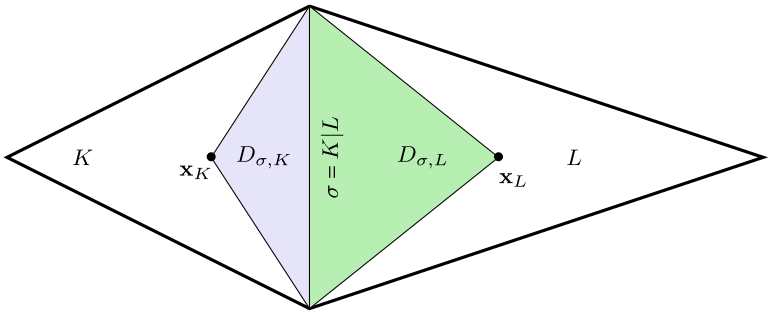


Fig. 1 Dual grid

$$v^{\text{out}}(x) = \lim_{\delta \to 0+} v(x + \delta \boldsymbol{n}), \qquad v^{\text{in}}(x) = \lim_{\delta \to 0+} v(x - \delta \boldsymbol{n}),$$
$$\overline{v}(x) = \frac{v^{\text{in}}(x) + v^{\text{out}}(x)}{2}, \qquad [\![v]\!] = v^{\text{out}}(x) - v^{\text{in}}(x).$$

Upwind Flux. Given a velocity $\boldsymbol{u}_h \in Q_h$ and a function $r_h \in Q_h$, we define for each face $\sigma \in \Sigma_{int}$ the upwind flux

$$\begin{aligned} Up[r_h, \boldsymbol{u}_h] &= r_h^{\text{up}} \boldsymbol{u}_h \cdot \boldsymbol{n} = r_h^{\text{in}} [\overline{\boldsymbol{u}_h} \cdot \boldsymbol{n}]^+ + r_h^{\text{out}} [\overline{\boldsymbol{u}_h} \cdot \boldsymbol{n}]^- \\ &= \overline{r_h}\, \overline{\boldsymbol{u}_h} \cdot \boldsymbol{n} - \frac{1}{2} |\overline{\boldsymbol{u}_h} \cdot \boldsymbol{n}| [\![r_h]\!], \end{aligned}$$

where

$$[f]^{\pm} = \frac{f \pm |f|}{2} \quad \text{and} \quad r^{\text{up}} = \begin{cases} r^{\text{in}} & \text{if } \overline{\boldsymbol{u}_h} \cdot \boldsymbol{n} \geq 0, \\ r^{\text{out}} & \text{if } \overline{\boldsymbol{u}_h} \cdot \boldsymbol{n} < 0. \end{cases}$$

Furthermore, we define the numerical flux function

$$F_h(r_h, \boldsymbol{u}_h) = Up[r_h, \boldsymbol{u}_h] - h^{\beta} [\![r_h]\!], \ 0 < \beta < 1.$$

Discrete Operators. For any $r_h, \boldsymbol{v}_h \in Q_h$ and $\boldsymbol{q}_h \in W_h$ we define the following discrete gradient and Laplace operators

$$\nabla_{\mathcal{D}} : Q_h \to W_h$$
$$\nabla_{\mathcal{D}} r_h = \sum_{\sigma \in \Sigma} (\nabla_{\mathcal{D}} r_h)_\sigma 1_{D_\sigma}, \quad (\nabla_{\mathcal{D}} r_h)_\sigma = \frac{1}{d_\sigma} [\![r_h]\!] \boldsymbol{n},$$
$$\nabla_h : Q_h \to Q_h$$
$$\nabla_h r_h = \sum_{K \in \mathcal{T}_h} (\nabla_h r_h)_K 1_K, \quad (\nabla_h r_h)_K = \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} \overline{r_h} \boldsymbol{n},$$
$$\Delta_h : Q_h \to Q_h$$
$$\Delta_h r_h = \sum_{K \in \mathcal{T}_h} (\Delta_h r_h)_K 1_K, \quad (\Delta_h r_h)_K = \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} \frac{[\![r_h]\!]}{d_\sigma},$$

and discrete divergence operators

$$\mathrm{div}_{\mathcal{T}} : W_h \to Q_h$$
$$\mathrm{div}_{\mathcal{T}} \boldsymbol{q}_h = \sum_{K \in \mathcal{T}_h} (\mathrm{div}_{\mathcal{T}} \boldsymbol{q}_h)_K 1_K, \quad (\mathrm{div}_{\mathcal{T}} \boldsymbol{q}_h)_K = \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} \boldsymbol{q}_h \cdot \boldsymbol{n},$$
$$\mathrm{div}_h : Q_h \to Q_h$$
$$\mathrm{div}_h \boldsymbol{v}_h = \sum_{K \in \mathcal{T}_h} (\mathrm{div}_h \boldsymbol{v}_h)_K 1_K, \quad (\mathrm{div}_h \boldsymbol{v}_h)_K = \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} \overline{\boldsymbol{v}_h} \cdot \boldsymbol{n},$$
$$\mathrm{div}_h^{\mathrm{up}} : Q_h \to Q_h$$
$$\mathrm{div}_h^{\mathrm{up}}(r_h \boldsymbol{v}_h) = \sum_{K \in \mathcal{T}_h} 1_K \mathrm{div}_h^{\mathrm{up}}(r_h \boldsymbol{v}_h)_K, \quad \mathrm{div}_h^{\mathrm{up}}(r_h \boldsymbol{v}_h)_K = \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} F_h(r_h, \boldsymbol{v}_h).$$

Further, the discrete symmetric gradient operator is given by

$$D_h(\boldsymbol{v}_h) = \frac{1}{2}(\nabla_h \boldsymbol{v}_h + \nabla_h \boldsymbol{v}_h^T), \quad \boldsymbol{v}_h \in Q_h.$$

Note that the operators $\nabla_{\mathcal{D}}$ and Δ_h can be extended to vector-valued functions componentwisely. Let $\boldsymbol{v}_h = (v_{1,h}, \ldots, v_{d,h}) \in Q_h$. Then we have

$$\nabla_{\mathcal{D}} \boldsymbol{v}_h = \left(\nabla_{\mathcal{D}} v_{1,h}, \ldots, \nabla_{\mathcal{D}} v_{d,h}\right), \quad \Delta_h \boldsymbol{v}_h = \left(\Delta_h v_{1,h}, \ldots, \Delta_h v_{d,h}\right),$$

and

$$(\nabla_{\mathcal{D}} \boldsymbol{v}_h)_\sigma = \frac{1}{d_\sigma} [\![\boldsymbol{v}_h]\!] \otimes \boldsymbol{n}, \quad (\Delta_h \boldsymbol{v}_h)_K = \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} \frac{[\![\boldsymbol{v}_h]\!]}{d_\sigma}.$$

2.2 Numerical Scheme for the Euler System

We recall a semi-discrete finite volume scheme for the Euler system Eq. 1,

$$\begin{aligned} D_t \varrho_h + \mathrm{div}_h^{\mathrm{up}}(\varrho_h \boldsymbol{u}_h) &= 0, \\ D_t \boldsymbol{m}_h + \mathrm{div}_h^{\mathrm{up}} F_h(\boldsymbol{m}_h, \boldsymbol{u}_h) + \nabla_h p_h &= h^{\alpha-1} \Delta_h \boldsymbol{u}_h, \\ D_t E_h + \mathrm{div}_h^{\mathrm{up}} F_h[E_h, \boldsymbol{u}_h] + \boldsymbol{u}_h \cdot \nabla_h p_h + p_h \mathrm{div}_h \boldsymbol{u}_h &= \frac{h^{\alpha-1}}{2} \Delta_h(\boldsymbol{u}_h^2), \end{aligned}$$

where $\boldsymbol{u}_h = \frac{\boldsymbol{m}_h}{\varrho_h}$, $p_h = (\gamma - 1)\left(E_h - \frac{1}{2}\frac{|\boldsymbol{m}_h|^2}{\varrho_h}\right)$ and D_t stands for the time derivative. The scheme was firstly introduced and studied in its weak form in our recent work [9]. Hereafter we will refer to it as the *FLM method*.

Definition 1 (FLM method) Given the initial values $(\varrho_{0,h}, \boldsymbol{m}_{0,h}, E_{0,h}) \in Q_h \times Q_h \times Q_h$, we seek a piecewise constant approximation $(\varrho_h, \boldsymbol{m}_h, E_h) \in Q_h \times Q_h \times Q_h$ which solves at any time $t \in (0, T]$ the following equations:

$$\int_\Omega D_t \varrho_h \phi_h \, \mathrm{d}x - \sum_{\sigma \in \Sigma_{int}} \int_\sigma F_h(\varrho_h, \boldsymbol{u}_h) [\![\phi_h]\!] \mathrm{d}S_x = 0, \ \forall\, \phi_h \in Q_h, \tag{5a}$$

$$\begin{aligned} \int_\Omega D_t \boldsymbol{m}_h \cdot \boldsymbol{\phi}_h \, \mathrm{d}x - \sum_{\sigma \in \Sigma_{int}} \int_\sigma \mathbf{F}_h(\boldsymbol{m}_h, \boldsymbol{u}_h) \cdot [\![\boldsymbol{\phi}_h]\!] \mathrm{d}S_x - \sum_{\sigma \in \Sigma_{int}} \int_\sigma \overline{p_h} \boldsymbol{n} \cdot [\![\boldsymbol{\phi}_h]\!] \mathrm{d}S_x \\ = -h^{\alpha-1} \sum_{\sigma \in \Sigma_{int}} \int_\sigma [\![\boldsymbol{u}_h]\!] \cdot [\![\boldsymbol{\phi}_h]\!] \mathrm{d}S_x, \ \forall\, \boldsymbol{\phi}_h \in Q_h, \end{aligned} \tag{5b}$$

$$\begin{aligned} &\int_\Omega D_t E_h \phi_h \, \mathrm{d}x - \sum_{\sigma \in \Sigma_{int}} \int_\sigma F_h(E_h, \boldsymbol{u}_h) [\![\phi_h]\!] \mathrm{d}S_x - \sum_{\sigma \in \Sigma_{int}} \int_\sigma \overline{p_h} [\![\phi_h \boldsymbol{u}_h]\!] \cdot \boldsymbol{n} \mathrm{d}S_x \\ &+ \sum_{\sigma \in \Sigma_{int}} \int_\sigma \overline{p_h \phi_h} [\![\boldsymbol{u}_h]\!] \cdot \boldsymbol{n} \mathrm{d}S_x = -\frac{h^{\alpha-1}}{2} \sum_{\sigma \in \Sigma_{int}} \int_\sigma [\![\boldsymbol{u}_h^2]\!] [\![\phi_h]\!] \mathrm{d}S_x, \ \forall\, \phi_h \in Q_h. \end{aligned} \tag{5c}$$

The initial values can be obtained by a standard projection onto the space Q_h,

$$\Pi_h[r]_{|_K} = \frac{1}{|K|} \int_K r \, \mathrm{d}x \ \text{ for any } K \in \mathcal{T}_h,$$

i.e. $(\varrho_{0,h}, \boldsymbol{m}_{0,h}, E_{0,h}) = (\Pi_h[\varrho_0], \Pi_h[\boldsymbol{m}_0], \Pi_h[E_0])$.

Remark 1 The FLM method in Eq. 5 can be also rewritten in the following per-cell flux formulation

$$D_t \varrho_K + \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} F_h(\varrho_h, \boldsymbol{u}_h) = 0,$$

$$D_t \boldsymbol{m}_K + \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} \left(\mathbf{F}_h(\boldsymbol{m}_h, \boldsymbol{u}_h) + \overline{p_h}\boldsymbol{n}\right) = h^{\alpha-1} \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} [\![\boldsymbol{u}_h]\!],$$

$$D_t E_K + \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} \left(F_h(E_h, \boldsymbol{u}_h) + (\overline{p_h}\boldsymbol{u}_h + p_h \overline{\boldsymbol{u}_h}) \cdot \boldsymbol{n} \right) = \frac{h^{\alpha-1}}{2} \sum_{\sigma \in \partial K} \frac{|\sigma|}{|K|} [\![\boldsymbol{u}_h^2]\!],$$

for any $K \in \mathcal{T}_h$.

Remark 2 Our finite volume method is based on an upwinding, which naturally yields a numerical diffusion. In addition we include a numerical diffusion of the form $h^{\beta+1}\Delta_h r_h$, where $r_h = \varrho_h, \boldsymbol{m}_h, E_h$. Altogether this diffusion is of the form $(\frac{1}{2}|\boldsymbol{u}_h \cdot \boldsymbol{n}| + h^\beta)h\Delta_h r_h$. Note that we have an additional numerical diffusion term $h^\alpha \Delta_h \boldsymbol{u}_h$ and $\frac{1}{2}h^\alpha \Delta_h \boldsymbol{u}_h^2$ in the momentum and energy equation, respectively. In the case the sound speed is larger than h^β, the numerical diffusion w.r.t. $\Delta_h r_h$ is smaller than that of standard numerical fluxes based on the Riemann problem solution.

It is truth, that we do not take a special care for the approximation of the contacts. On the other hand, a better resolution can be achieved by introducing a linear reconstruction and limiting to obtained second-order extension.

2.2.1 Properties of the FLM Method

For the rigorous convergence analysis of scheme in Eq. 5 a few important properties are inevitable.

- **Existence of numerical solution**
  The discrete problem Eq. 5 admits a solution $(\varrho_h(t), \boldsymbol{m}_h(t), E_h(t)) \in Q_h \times Q_h \times Q_h$, for any $t \geq 0$. As shown in [9], the result follows from the standard theory of ODEs and sufficiently strong *a priori* bounds.
- **Conservation of discrete mass and energy**
  In a straightforward way it can be shown that

$$\int_\Omega \varrho_h(t, \cdot)\, \mathrm{d}x = \int_\Omega \varrho_{0,h}\, \mathrm{d}x = \tilde{M}_0 > 0,$$
$$\int_\Omega E_h(t, \cdot)\, \mathrm{d}x = \int_\Omega E_{0,h}\, \mathrm{d}x = \tilde{E}_0 > 0,\ t \geq 0.$$

- **Positivity of the discrete density, pressure and temperature**
  For any fixed h, the approximate density, pressure and consequently also temperature remain strictly positive on any finite time interval. We refer the reader to [9, Sections 4.3, 4.4] for more details.

- **Discrete entropy inequality**
  The discrete (renormalized) entropy inequality in the sense of Tadmor is satisfied, cf. [20, 21]. More precisely, it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{T}_h}\varrho_h\chi(s_h)\Phi_h\,\mathrm{d}x \geq \sum_{\sigma\in\Sigma_{int}}\int_\sigma Up[\varrho_h\chi(s_h),\boldsymbol{u}_h][[\Phi_h]]\mathrm{d}S_x +$$
$$+\sum_{\sigma\in\Sigma_{int}}\int_\sigma \mu_h\left(\overline{\nabla_\varrho(\varrho_h\chi(s_h))}[[\varrho_h]]+\overline{\nabla_p(\varrho_h\chi(s_h))}[[p_h]]\right)[[\Phi_h]]\mathrm{d}S_x,$$

  where χ is a non-decreasing, concave, twice continuously differentiable function on R that is bounded from above. For the derivation and proof see [9, Section 3.2].
- **Minimum entropy principle**
  The discrete physical entropy $s_h = \log\left(\vartheta_h^{c_v}/\varrho_h\right)$ attains its minimum at the initial time, cf. [13, 22], i.e.,

$$s_h(t)\geq s_0,\ t\geq 0,\quad \text{where}\quad -\infty < s_0 < \min s_h(0).$$

  The entropy is either constant or produced over time, thus the second law of thermodynamics holds. See [9, Section 4.2] for more details.

Clearly, the FLM method belongs to the class of invariant domain preserving schemes introduced in [13, 14]. Based on the above properties the following convergence result for the FLM method was proved in [9].

Theorem 1 (Convergence of the FLM method) *Let the initial data* $(\varrho_{0,h}, \boldsymbol{m}_{0,h}, E_{0,h})$ *satisfy*

$$\varrho_{0,h}\geq\underline{\varrho}>0,\ E_{0,h}-\frac{1}{2}\frac{|\boldsymbol{m}_{0,h}|^2}{\varrho_{0,h}}>0.$$

Let $(\varrho_h, \boldsymbol{m}_h, E_h)\in Q_h\times Q_h\times Q_h$ *be the solution of the scheme Eq. 5 such that*

$$0<\beta<1,\ 0<\alpha<\frac{4}{3},$$

and

$$0<\underline{\varrho}\leq\varrho_h(t),\ \vartheta_h(t)\leq\overline{\vartheta}\ \textit{for all}\ t\in[0,T]\ \textit{uniformly for}\ h\to 0.$$

Then the family of approximate solutions $\{\varrho_h, \boldsymbol{m}_h, E_h\}_{h>0}$ *generates a dissipative measure–valued (DMV) solution of the complete Euler system Eq. 1 in the sense of [2].*

Let us point out that a DMV solution of the Euler system is a time-space parametrized probability measure, i.e. the Young measure. The expected values of density and entropy with respect to this Young measure satisfy the corresponding weak formulation of mass conservation and entropy inequality, respectively. The weak formulation

for the expected value of the momentum allows a concentration defect that is controlled by the dissipation in the energy balance. The energy conservation is relaxed and the expected value of the energy dissipates in time, see [2] and [9].

Furthermore, evoking the DMV–strong uniqueness result proved in [2, Theorem 3.3] we obtain the following strong convergence result.

Theorem 2 (Strong convergence of the FLM method) *In addition to the hypotheses of Theorem 1, suppose that the complete Euler system Eq. 1 admits a Lipschitz–continuous solution* $(\varrho, \boldsymbol{m}, E)$ *defined on* $[0, T]$.

Then

$$\varrho_h \to \varrho,\ \boldsymbol{m}_h \to \boldsymbol{m},\ E_h \to E \text{ (strongly) in } L^1((0,T)\times\Omega).$$

In Sect. 3 we will illustrate numerical behavior of the FLM method on a series of well-known benchmarks. In what follows we recall the extension of the FLM method to the finite volume method for the Navier–Stokes–Fourier system introduced in [11]. It turned out that for the convergence analysis of the latter system it is more convenient to work with the temperature formulation instead of the internal energy in the last equation of Eq. 2.

2.3 Numerical Scheme for the Navier–Stokes–Fourier System

Having introduced the notation in Sect. 2.1, we now present a semi-discrete finite volume approximation of the Navier–Stokes–Fourier (NSF) system Eq. 2,

$$\begin{aligned} D_t \varrho_h + \mathrm{div}_h^{\mathrm{up}}(\varrho_h \boldsymbol{u}_h) &= 0,\\ D_t(\varrho_h \boldsymbol{u}_h) + \mathrm{div}_h^{\mathrm{up}}(\varrho_h \boldsymbol{u}_h, \boldsymbol{u}_h) + \nabla_h p_h &= 2\mu \mathrm{div}_h D_h(\boldsymbol{u}_h) + \lambda \nabla_h \mathrm{div}_h \boldsymbol{u}_h,\\ c_v D_t(\varrho_h \vartheta_h) + c_v \mathrm{div}_h^{\mathrm{up}}(\varrho_h \vartheta_h, \boldsymbol{u}_h) - \kappa \Delta_h \vartheta_h &\\ = 2\mu\, |\mathbf{D}_h(\boldsymbol{u}_h)|^2 + \lambda\, |\mathrm{div}_h \boldsymbol{u}_h|^2 - p_h \mathrm{div}_h \boldsymbol{u}_h.& \end{aligned}$$

Note that a fully discrete (implicit in time) version of this scheme was analyzed in our work [11].

Definition 2 (Finite volume (FV) method for NSF) Given the initial values $(\varrho_{0,h}, \boldsymbol{u}_{0,h}, \vartheta_{0,h}) \in Q_h \times Q_h \times Q_h$, we seek a piecewise constant approximation $(\varrho_h, \boldsymbol{u}_h, \vartheta_h) \in Q_h \times Q_h \times Q_h$ which solves at any time $t \in (0, T]$ the following equations:

$$\int_\Omega D_t \varrho_h\, \phi_h \,\mathrm{d}x - \sum_{\sigma \in \Sigma_{int}} \int_\sigma F_h(\varrho_h, \boldsymbol{u}_h) [\![\phi_h]\!] \mathrm{d}S_x = 0,\ \forall \phi_h \in Q_h, \tag{6a}$$

$$\int_\Omega D_t(\varrho_h \boldsymbol{u}_h)\cdot\boldsymbol{\phi}_h\,\mathrm{d}x - \sum_{\sigma\in\Sigma_{int}}\int_\sigma \mathbf{F}_h(\varrho_h\boldsymbol{u}_h,\boldsymbol{u}_h)\cdot[\![\boldsymbol{\phi}_h]\!]\mathrm{d}S_x - \int_\Omega p_h \mathrm{div}_h\boldsymbol{\phi}_h\,\mathrm{d}x$$
$$= -2\mu\int_\Omega D_h(\boldsymbol{u}_h):\nabla_h\boldsymbol{\phi}_h\,\mathrm{d}x - \lambda\int_\Omega \mathrm{div}_h\boldsymbol{u}_h\mathrm{div}_h\boldsymbol{\phi}_h\,\mathrm{d}x,\ \forall\boldsymbol{\phi}_h\in Q_h, \quad (6b)$$

$$c_v\int_\Omega D_t(\varrho_h\vartheta_h)\,\phi_h\,\mathrm{d}x - c_v\sum_{\sigma\in\Sigma_{int}}\int_\sigma F_h(\varrho_h\vartheta_h,\boldsymbol{u}_h)[\![\phi_h]\!]\mathrm{d}S_x - \kappa\int_\Omega \Delta_h\vartheta_h\,\phi_h\,\mathrm{d}x$$
$$= \int_\Omega \big(2\mu\,|\mathbf{D}_h(\boldsymbol{u}_h)|^2 + \lambda\,|\mathrm{div}_h\boldsymbol{u}_h|^2 - p_h\mathrm{div}_h\boldsymbol{u}_h\big)\phi_h\,\mathrm{d}x,\ \forall\phi_h\in Q_h. \quad (6c)$$

Remark 3 Let us point out that the $h^{\alpha-1}$-terms in Eq. 5b and Eq. 5c yield an additional diffusion and make the FLM method a particular vanishing viscosity approximation of the Euler system. Since the physical viscosity is naturally included in the Navier–Stokes–Fourier system, we do not need to include the additional diffusion in Eq. 6.

Remark 4 The numerical scheme in Eq. 6 can be also rewritten in the usual finite volume formulation for any $K\in\mathcal{T}_h$,

$$D_t\varrho_K + \sum_{\sigma\in\partial K}\frac{|\sigma|}{|K|}F_h(\varrho_h,\boldsymbol{u}_h) = 0,$$
$$D_t(\varrho\boldsymbol{u})_K + \sum_{\sigma\in\partial K}\frac{|\sigma|}{|K|}\left(\mathbf{F}_h(\varrho_h\boldsymbol{u}_h,\boldsymbol{u}_h) + \overline{p_h}\boldsymbol{n}\right)$$
$$= \sum_{\sigma\in\partial K}\frac{|\sigma|}{|K|}\left(2\mu\overline{D_h(\boldsymbol{u}_h)}\cdot\boldsymbol{n} + \lambda\overline{\mathrm{div}_h\boldsymbol{u}_h}\boldsymbol{n}\right),$$
$$c_v D_t(\varrho\vartheta)_K + \sum_{\sigma\in\partial K}\frac{|\sigma|}{|K|}\left(c_v F_h(\varrho_h\vartheta_h,\boldsymbol{u}_h) - \kappa\frac{[\![\vartheta_h]\!]}{d_\sigma}\right)$$
$$= \sum_{\sigma\in\partial K}\frac{|\sigma|}{|K|}\left(2\mu\,|\mathbf{D}_h(\boldsymbol{u}_h)|_K^2 + \lambda\,|\mathrm{div}_h\boldsymbol{u}_h|_K^2 - p_K(\mathrm{div}_h\boldsymbol{u}_h)_K\right).$$

2.3.1 Properties of the FV Method for NSF

Analogously as in the inviscid case for the convergence analysis it is fundamental that our numerical scheme fulfills some invariant domain preserving properties. In [11] we have proved the following:

- **Conservation of discrete mass**
  One can easily show that

$$\int_\Omega \varrho_h(t,\cdot)\,\mathrm{d}x = \int_\Omega \varrho_{0,h}\,\mathrm{d}x = \tilde{M}_0 > 0,\ \ t \geq 0.$$

- **Non-negativity of the discrete density**
  The approximate density remains non-negative on any finite time interval.
- **Discrete total energy dissipation**
  Let $(\varrho_h, \boldsymbol{u}_h, \vartheta_h) \in Q_h \times Q_h \times Q_h$ be a solution to Eq. 6. Then

$$E_h(t) \leq E_0,\ \ t \geq 0,$$

  where

$$E_h(t) = \int_\Omega \left(\frac{1}{2}\varrho_h(t)|\boldsymbol{u}_h(t)|^2 + c_v \varrho_h(t)\vartheta_h(t) \right) \mathrm{d}x.$$

  See [11, Theorem 3.1] for the proof.
- **Discrete entropy inequality**
  The scheme in Eq. 6 is entropy stable. It holds that

$$\int_\Omega D_t\left(\varrho_h s_h\right)\,\mathrm{d}x \geq -\int_\Omega \kappa \nabla_{\mathcal{D}}\vartheta_h \cdot \nabla_{\mathcal{D}}\left(\frac{1}{\vartheta_h}\right)\mathrm{d}x + \int_\Omega \frac{1}{\vartheta_h}\left(2\mu|\mathbf{D}(\boldsymbol{u}_h)|^2 + \lambda|\mathrm{div}_h \boldsymbol{u}_h|^2\right)\mathrm{d}x,$$

  see [11, Lemma 3.4].

Remark 5 Note that the above properties shown in [11] for a fully discrete implicit in time version of scheme Eq. 6 can be proven in a straightforward manner for the semi-discrete scheme presented here.

The structure preserving properties listed above, together with the assumptions on uniform boundedness of the discrete density and temperature, are sufficient to derive suitable *a priori* estimates and consistency formulation of scheme Eq. 6 which are inevitable for the convergence of its solutions. We now recall the convergence results proved in [11].

Theorem 3 (Convergence of the FV method for NSF) *Let the initial data satisfy the assumptions*

$$0 < \underline{\varrho} \leq \varrho_{0,h} \leq \overline{\varrho},\ \ 0 < \underline{\vartheta} \leq \vartheta_{0,h} \leq \overline{\vartheta},\ \ \|\boldsymbol{u}_{0,h}\|_{L^2} \leq \overline{u},$$

for some positive constants $\underline{\varrho}, \overline{\varrho}, \underline{\vartheta}, \overline{\vartheta}, \overline{u}$. *Let* $(\varrho_h, \vartheta_h, \boldsymbol{u}_h) \in Q_h \times Q_h \times Q_h$ *be the solution of the finite volume scheme Eq. 6, satisfying the assumptions*

$$0 < \underline{\varrho} \leq \varrho_h(t) \leq \overline{\varrho}, 0 < \underline{\vartheta} \leq \vartheta_h(t) \leq \overline{\vartheta} \text{ uniformly for } h \to 0, \text{and all } t \in (0, T).$$

Then the family $\{\varrho_h, \vartheta_h, \boldsymbol{u}_h, \boldsymbol{D}_h(\boldsymbol{u}_h), \nabla_{\mathcal{D}}\vartheta_h\}_{h>0}$ *generates a DMV solution of the Navier–Stokes–Fourier system Eq. 2 in the sense of [3].*

Analogously as for the inviscid flows a DMV solution is the Young measure. Expected values of density, momentum, energy and entropy satisfy appropriate generalized formulation of Eq. 2. Further, applying the DMV–strong uniqueness principle established in [3, Theorem 6.1] and [11, Theorem 5.5] we have the following strong convergence result.

Theorem 4 (Strong convergence of the FV method for NSF) *In addition to the hypotheses of Theorem 3 assume that* $\{\mathcal{V}_{t,x}\}_{(t,x)\in(0,T)\times\Omega}$ *is a DMV solution of the Navier–Stokes–Fourier system Eq. 2 in the sense of [3] such that*

$$\mathcal{V}_{t,x}\left\{0 < \underline{\varrho} \leq \varrho \leq \overline{\varrho},\ \vartheta \leq \overline{\vartheta},\ |\boldsymbol{u}| \leq \overline{u}\right\} = 1 \text{ for a.a. } (t,x) \in (0,T)\times\Omega \tag{7}$$

for some constants $\underline{\varrho}$, $\overline{\varrho}$, $\overline{\vartheta}$, *and* $\overline{u}$. *Let, moreover,*

$$\mathcal{V}_{0,x} = \delta_{\varrho_0(x),\vartheta_0(x),\boldsymbol{u}_0(x)} \quad \text{for a.a. } x \in \Omega,$$

where $(\varrho_0, \vartheta_0, \boldsymbol{u}_0)$ *belongs to the regularity class*

$$\varrho_0, \vartheta_0 \in W^{3,2}(\Omega),\ \varrho_0,\ \vartheta_0 > 0 \text{ in } \Omega,\ \boldsymbol{u}_0 \in W_0^{3,2}(\Omega; R^3). \tag{8}$$

Finally, suppose that the Navier–Stokes–Fourier system Eq. 2 is endowed with the initial data $(\varrho_0, \vartheta_0, \boldsymbol{u}_0)$ *satisfying Eq. 8. Let* $(\varrho_h, \vartheta_h, \boldsymbol{u}_h)$ *be the solution of the finite volume scheme Eq. 6, and in addition,*

$$|\boldsymbol{u}_h(t)| \leq \overline{u} \text{ uniformly for } h \to 0 \text{ and all } t \in (0,T).$$

Then

$$\begin{aligned}
&\varrho_h \to \varrho \text{ (strongly) in } L^p((0,T)\times\Omega),\\
&\vartheta_h \to \vartheta \text{ (strongly) in } L^p((0,T)\times\Omega),\\
&\boldsymbol{u}_h \to \boldsymbol{u} \text{ (strongly) in } L^p\left((0,T)\times\Omega; R^d\right),\ p \in [1,\infty),
\end{aligned}$$

where $(\varrho, \vartheta, \boldsymbol{u})$ *is a strong (classical) solution of the Navier–Stokes–Fourier system.*

3 Numerical Experiments

In this section we demonstrate the performance of both finite volume methods, the FLM method, see Eq. 5, for the Euler equations, and the finite volume method, see Eq. 6, for the Navier–Stokes–Fourier equations.

For time discretization we use the forward finite differences which yield the explicit finite volume scheme for the Euler system. Diffusive fluxes in the Navier–Stokes–Fourier equations are approximated by the backward finite differences and thus implicitly in time. For stability reasons, we set the time step as $\Delta t = \min\{\Delta t_a, \Delta t_b\}$ in each sub-iteration. The first term arises from the CFL stability condition: $\Delta t_a = \mathrm{CFL}\, h / \max\{|\boldsymbol{u}| + c\}$, $c = \sqrt{\vartheta}$. In our numerical simulations we set $\mathrm{CFL} = 0.5$ if not explicitly claimed otherwise. The second term is due to the parabolic regularization: $\Delta t_b = h^{1-\beta}/(2d)$.

3.1 Numerical Experiments for the FLM Method

3.1.1 Experimental Order of Convergence (EOC)

We aim to validate the theoretical result on the convergence of ϱ, $\boldsymbol{m}$, E presented in Theorem 2 by computing the corresponding norms of numerical errors

$$\left\| e_f \right\| = \frac{\left\| f - f_{ref} \right\|_{L_t^1 L_x^1}}{\left\| f_{ref} \right\|_{L_t^1 L_x^1}}, \quad f \in \{\varrho, \boldsymbol{m}, E\},$$

where $L_t^1 L_x^1$ is a shortening for $L^1(0, T; L^1(\Omega))$. Analogous notation is used for other Bochner spaces below. Additionally, we also provide the numerical errors of the velocity $\boldsymbol{u}$ in $L_t^2 L_x^2-$norm and pressure p in $L_t^\infty L_x^1-$norm. The reference solution is the exact solution to Eq. 1

$$\begin{aligned} \varrho_{ref} &= 2 + \cos(2\pi x), \quad \boldsymbol{u}_{ref} = \frac{\sin(\pi t)}{2 + \cos(2\pi x)} \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \\ p_{ref} &= (2 + \cos(2\pi x))(2 + \sin(2\pi x)), \qquad x \in [0, 1]. \end{aligned} \tag{9}$$

Setting $\gamma = 1.4$, $\alpha = 1.3$, $\beta = 0.2$ and $\mathrm{CFL} = 0.6$, we observe the first order convergence rate for the FLM method, see Table 1.

Table 1 Relative errors and EOC for the FLM method at time $t = 0.1$

h	$\|\|e_\varrho\|\|$	EOC	$\|\|e_{\boldsymbol{m}}\|\|$	EOC	$\|\|e_E\|\|$	EOC	$\|\|e_{\boldsymbol{u}}\|\|$	EOC	$\|\|e_p\|\|$	EOC
1/32	9.00e−03	–	4.15e−02	–	1.21e−02	–	5.75e−02	–	1.94e−02	–
1/64	4.05e−03	1.15	1.88e−02	1.14	5.40e−03	1.16	2.65e−02	1.12	8.74e−03	1.15
1/128	1.81e−03	1.16	8.36e−03	1.17	2.41e−03	1.16	1.20e−02	1.14	3.94e−03	1.15
1/256	8.07e−04	1.17	3.71e−03	1.17	1.08e−03	1.16	5.41e−03	1.15	1.78e−03	1.15

Table 2 Initial data of 1D tests

Test	ϱ_L	u_L	p_L	ϱ_R	u_R	p_R	T_{max}	x_m
1	1.0	−2.0	0.4	1.0	2.0	0.4	0.15	0.5
2	1.0	0.0	1000.0	1.0	0.0	0.01	0.012	0.5
3	1.4	0.0	1.0	1.0	0.0	1.0	2.0	0.5
4	1.4	0.1	1.0	1.0	0.1	1.0	2.0	0.5

3.1.2 1D Benchmark Problems

We test one-dimensional Riemann problems studied in [15, 23] with the initial data

$$(\varrho, u, p) = \begin{cases} (\varrho_L, u_L, p_L) & \text{if } 0 \le x < x_m, \\ (\varrho_R, u_R, p_R) & \text{if } x_m \le x \le 1, \end{cases}$$

with the corresponding values presented in Table 2.

Test 1 has a weak solution consisting of two rarefaction waves and it is typically used for checking the positivity of density; Test 2 is designed for strong shock; Test 3 and 4 are designed to capture stationary contact waves. We set $\gamma = 1.4$, $\beta = 0.2$ and aim to show the numerical performance of the scheme Eq. 5 on the domain $\Omega = [0, 1]$ with mesh size $h = 1/400$. First, we present in Fig. 2 the results of numerical simulations for different choices of α, that is the parameter appearing in the artificial diffusion terms in Eq. 5b and Eq. 5c. Secondly, we show in Fig. 3 the comparison of the numerical solutions obtained by the FLM method with that of the HLL finite volume method [23].

3.1.3 2D Benchmark Problems

Now we test the two-dimensional Riemann problems studied in [15–17] with $\Omega = [-1, 1]^2$. Boundary values are obtained by extrapolation of conservative variables $(\varrho, \boldsymbol{m}, E)$.

Test 1: circular two-dimensional Sod problem with the initial data

$$(\varrho, u_1, u_2, p) = \begin{cases} (1.0, 0, 0, 1.0), & |x| < 0.4, \\ (1.0, 0, 0, 0.1), & \text{else.} \end{cases}$$

Figure 4 displays the contour lines of the numerical solution of density, velocity components, and pressure at time $t = 0.2$ which are in a very good agreement with the results presented in literature, cf., e.g., [23].

Test 2: two-dimensional benchmark Riemann problem consisting of two moving shocks and two standing slip lines. The initial values are set as

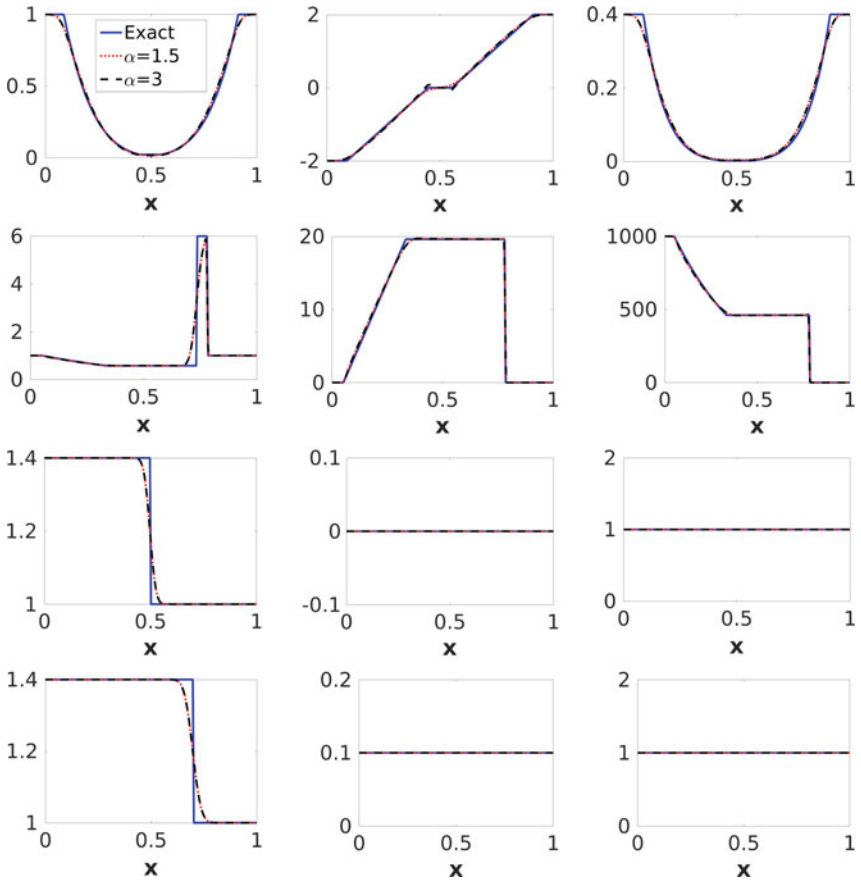


Fig. 2 1D tests: from top to bottom are Tests 1 to 4, from left to right numerical solutions of ϱ, $\boldsymbol{u}$, p. The solid blue lines represent solutions obtained by the exact Riemann solver. The dotted red lines and the dashed black lines are solutions obtained by the FLM scheme with $\alpha = 1.5$ and $\alpha = 3$, respectively

$$(\varrho, u_1, u_2, p) = \begin{cases} (0.5313, 0, 0.7276, 0.4), & x > 0, y > 0, \\ (1.0, 0.7276, 0, 1.0), & x < 0, y > 0, \\ (0.8, 0, 0, 1.0), & x < 0, y < 0, \\ (1.0, 0, 0.7276, 1.0), & x > 0, y < 0. \end{cases}$$

Figure 5 shows the numerical solution for density and pressure for different CFL numbers. Numerical solutions obtained by the FLM method are in good agreement with the results presented in literature, see, e.g., [16].

Test 3: two-dimensional Riemann problem with the initial condition

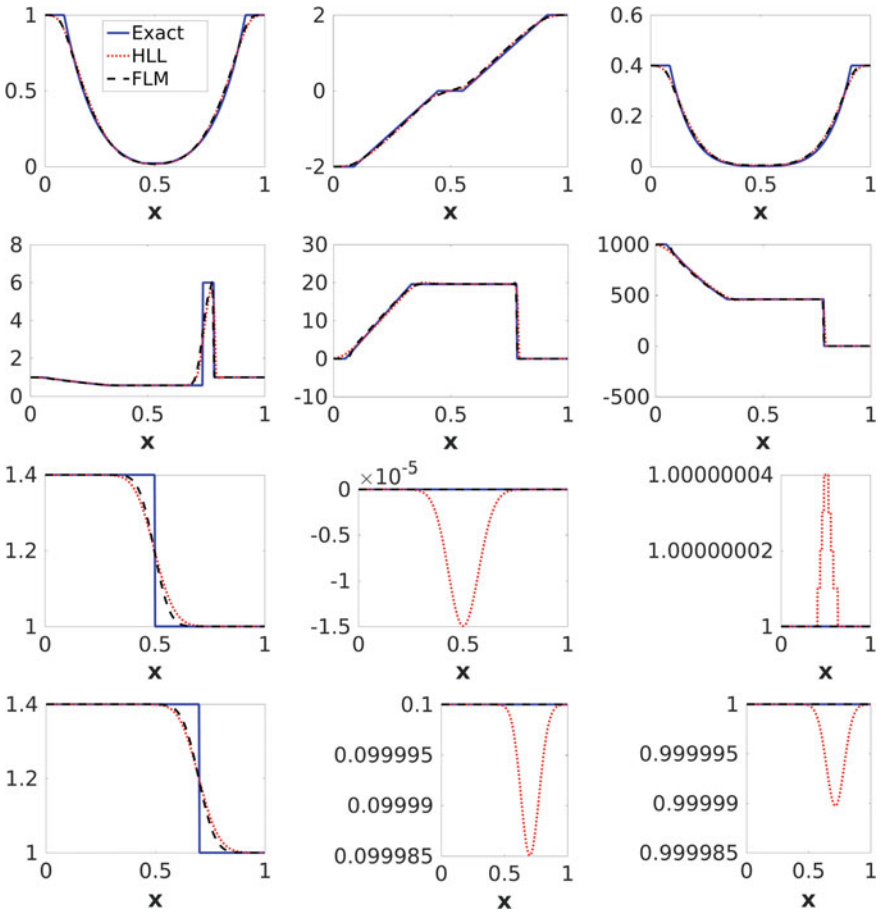
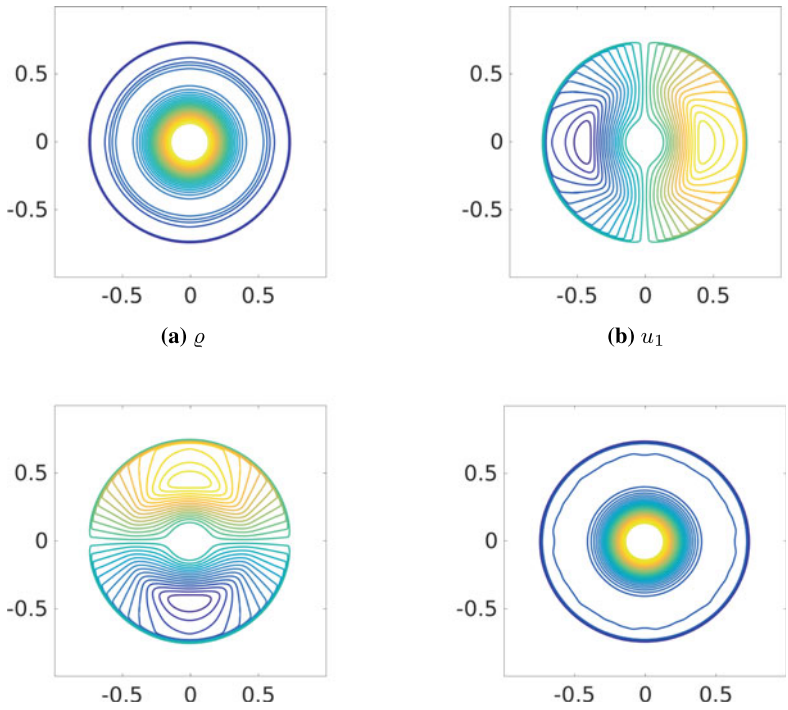


Fig. 3 1D tests: from top to bottom are Tests 1 to 4, from left to right numerical solutions of ϱ, $\boldsymbol{u}$, p. The solid blue lines, the dotted red lines and the dashed black lines are solutions obtained by the exact Riemann, HLL, and FLM ($\alpha = 1.5$) solvers, respectively

$$(\varrho, u_1, u_2, p) = \begin{cases} (1.1, 0, 0, 1.1), & x > 0, y > 0, \\ (0.5065, 0, 0.8939, 0.35), & x < 0, y > 0, \\ (1.1, 0.8939, 0.8939, 1.1), & x < 0, y < 0, \\ (0.5065, 0, 0.8939, 0.35), & x > 0, y < 0. \end{cases}$$

In this configuration there are two forward moving shocks and two backward moving shocks. Figure 6 depicts the contour lines of the numerical solution of density, velocity components, and pressure at time $t = 0.25$. We can again confirm that the numerical solution is in good agreement with the results presented in the literature, cf., e.g., [16].



(a) ϱ **(b)** u_1

(c) u_2 **(d)** p

Fig. 4 Test 1: Sod problem solution on rectangular mesh $h_x = h_y = 0.05$ with $\alpha = 1.5$, $\beta = 0.2$ at time $t = 0.2$

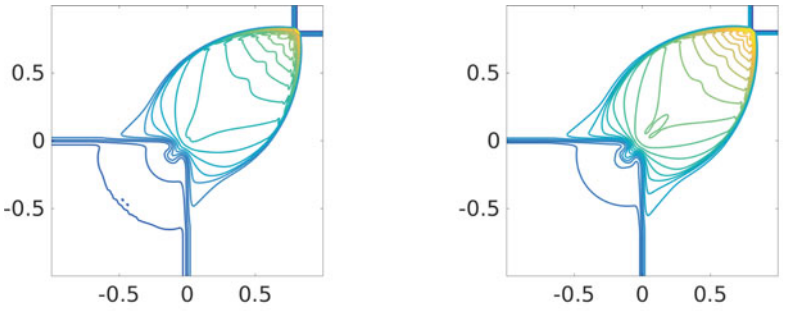
3.2 Numerical Experiments for the FV Method for NSF

3.2.1 Experimental Order of Convergence (EOC)

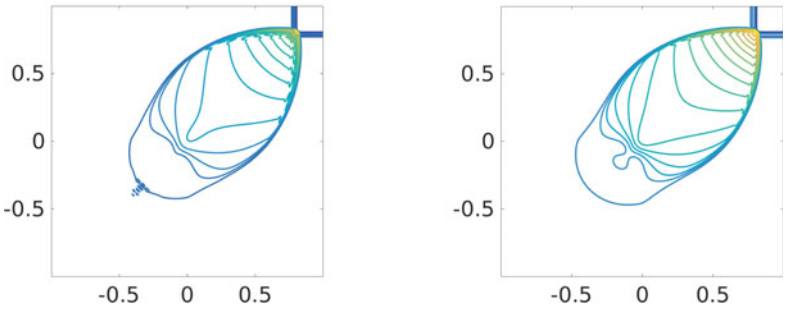
Our aim in this section is to validate theoretical results on the convergence of ϱ, $\boldsymbol{u}$, ϑ presented in Theorem 4 by computing the numerical errors

$$\|e_f\| = \frac{\|f - f_{ref}\|_{L_t^q L_x^q}}{\|f_{ref}\|_{L_t^q L_x^q}}, \quad f \in \{\varrho, \boldsymbol{u}, \vartheta\}, \ q = 1, 2.$$

Here the reference solution is the same as in Eq. 9. Thus, we have a manufactured exact solution with a suitable external force in the momentum and energy equation. Setting $\mu = \lambda = \kappa = 1$, $\beta = 0.2$ and CFL $= 0.6$ we observe the first order convergence rate for the scheme Eq. 6, see Table 3. We can observe the first order convergence on rectangular as well as triangular mesh.



(a) CFL = 0.6 **(b)** CFL = 0.3

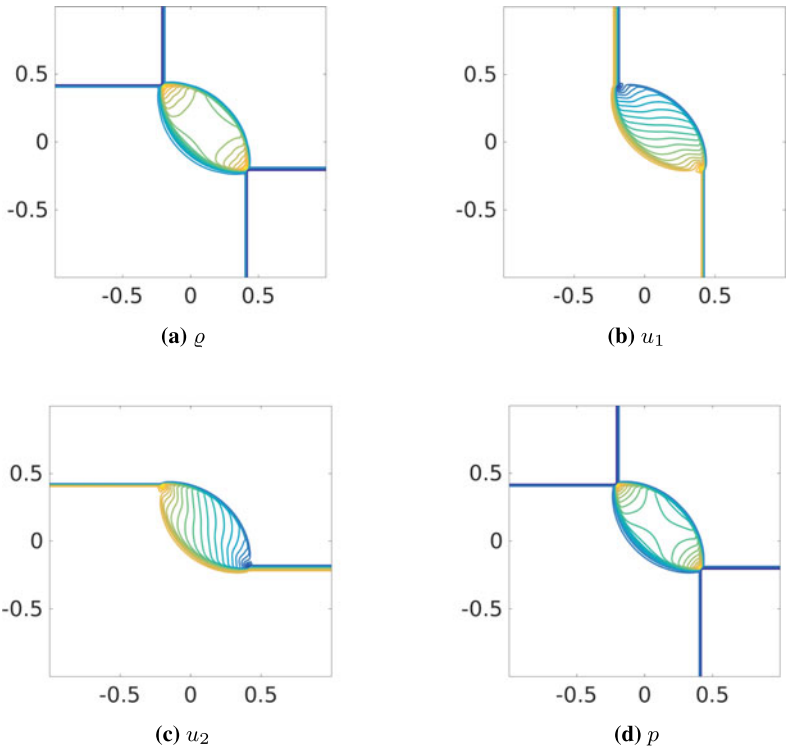


(c) CFL = 0.6 **(d)** CFL = 0.3

Fig. 5 Test 2: solution of ϱ (upper row) and p (lower row) on rectangular mesh $h_x = h_y = 0.05$ with $\alpha = 1.5$, $\beta = 0.5$ at time $t = 0.52$

Table 3 Relative errors and EOC for the FV method for NSF at time $t = 0.1$

h	$\|e_\varrho\|$	EOC	$\|e_{\boldsymbol{u}}\|$	EOC	$\|e_\vartheta\|$	EOC	$\|e_\varrho\|$	EOC	$\|e_{\boldsymbol{u}}\|$	EOC	$\|e_\vartheta\|$	EOC
	$L^1((0,T)\times\Omega)$-norm						$L^2((0,T)\times\Omega)$-norm					
Rectangular mesh												
32	2.09e−02	–	2.24e−02	–	1.27e−02	–	2.52e−02	–	2.71e−02	–	1.49e−02	–
64	9.51e−03	1.14	1.06e−02	1.08	5.78e−03	1.13	1.15e−02	1.12	1.31e−02	1.05	6.84e−03	1.12
128	4.27e−03	1.16	4.87e−03	1.12	2.60e−03	1.15	5.21e−03	1.15	6.10e−03	1.10	3.09e−03	1.15
256	1.90e−03	1.16	2.21e−03	1.14	1.16e−03	1.16	2.34e−03	1.16	2.80e−03	1.12	1.38e−03	1.16
512	8.49e−04	1.17	9.98e−04	1.15	5.20e−04	1.16	1.05e−03	1.16	1.27e−03	1.14	6.19e−04	1.16
Triangular mesh												
1/32	9.17e−03	–	1.23e−02	–	4.90e−03	–	1.10e−02	–	1.60e−02	–	5.82e−03	–
1/64	4.02e−03	1.19	6.68e−03	0.89	2.43e−03	1.01	4.83e−03	1.18	8.79e−03	0.87	2.92e−03	0.99
1/128	1.78e−03	1.18	4.10e−03	0.70	1.20e−03	1.02	2.13e−03	1.18	5.44e−03	0.69	1.45e−03	1.01
1/256	7.92e−04	1.17	2.99e−03	0.46	5.87e−04	1.03	9.50e−04	1.17	3.94e−03	0.46	7.19e−04	1.01
1/512	3.56e−04	1.15	2.53e−03	0.24	2.89e−04	1.02	4.27e−04	1.15	3.31e−03	0.25	3.57e−04	1.01



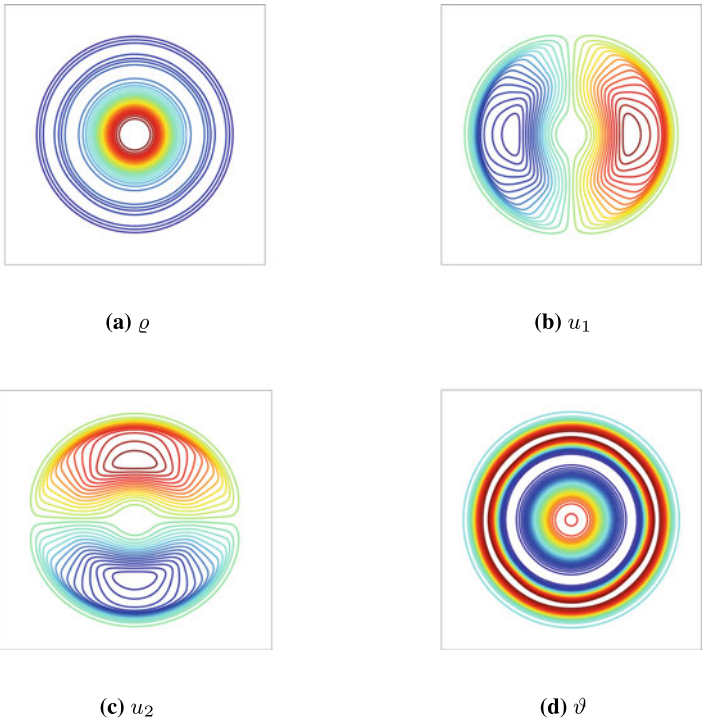
(a) ϱ **(b)** u_1

(c) u_2 **(d)** p

Fig. 6 Test 3: solution of ϱ, u_1, u_2, and p on rectangular mesh $h_x = h_y = 0.05$ at time $t = 0.25$

3.2.2 2D Benchmark Problems

Test 4: Circular shock problem.
We again test the two-dimensional Sod problem using the same initial data as in the first experiment of Sect. 3.1.3 with $\mu = \lambda = \kappa = 0.001$ and CFL $= \beta = 0.6$. The contour lines of the numerical solutions are shown in Fig. 7. Small viscosity effects can be noticed but overall the numerical solutions for inviscid and viscous case are similar as expected.



(a) ϱ **(b)** u_1

(c) u_2 **(d)** ϑ

Fig. 7 Test 4: Circular shock solution on rectangular mesh $h_x = h_y = 0.05$ at time $t = 0.2$

Test 5: Gresho Vortex problem with the initial data [15]

$$(u,\, p)(r) = \begin{cases} (5r,\ 5 + 12.5r^2) & r < 0.2, \\ (2 - 5r,\ 9 - 4\ln 0.2 + 12.5r^2 - 20r + 4\ln r) & 0.2 \le r < 0.4, \\ (0,\ 3 + 4\ln 2) & r > 0.4. \end{cases}$$

Figure 8 displays the contour lines of the numerical solutions obtained by the scheme Eq. 6 with the parameters $\mu = \lambda = \kappa = 0.01$, and CFL $= \beta = 0.6$ at time $t = 0.2$.

(a) ϱ **(b)** u_1

(c) u_2 **(d)** ϑ

Fig. 8 Test 5: Gresho vortex solution on rectangular mesh $h_x = h_y = 0.05$ at time $t = 0.2$

Conclusion

We have presented behavior and performance of two new convergent finite volume methods for compressible fluids, both inviscid and viscous. These new finite volume methods satisfy some important invariant domain preserving properties, such as the minimum entropy principle, mass and energy conservation, positivity preservation, total energy dissipation and entropy production. These are crucial for showing the stability and consistency of the schemes. In the framework of a nonlinear version of the Lax equivalence theorem, see [9, 11], these properties directly imply the strong convergence of numerical solutions to a strong solution on its lifespan. Our numerical experiments presented in Sect. 3 confirm these theoretical convergence results.

Acknowledgement M. Lukáčová-Medviďová has been partially supported by the German Science Foundation under the grants TRR 146 Multiscale simulation methods for soft matter systems and TRR 165 Waves to weather. H. Mizerová and B. She have received funding from the Czech Science Foundation (GAČR), Grant Agreement 18–05974S. The Institute of Mathematics of the Czech Academy of Sciences is supported by RVO:67985840.

References

1. Ben-Artzi, M., Li, J., Warnecke, G.: A direct Eulerian GRP scheme for compressible fluid flows. J. Comput. Phys. **218**(1), 19–43 (2006)
2. Březina, J., Feireisl, E.: Measure-valued solutions to the complete Euler system. J. Math. Soc. Jpn. **70**(4), 1227–1245 (2018)
3. Březina, J., Feireisl, E., Novotný, A.: Stability of strong solutions to the Navier-Stokes-Fourier system. SIAM J. Math. Anal. **52**(2), 1761–1785 (2020)
4. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (2002)
5. Cockburn, B., Shu, C.: TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework. Math. Comp. **52**(186), 411–435 (1989)
6. Dolejší, V., Feistauer, M.: Discontinuous Galerkin method: analysis and applications to compressible flow. Springer Series in Computational Mathematics, vol. 48, Springer-Verlag (2015)
7. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Handb. Numer. Anal. **7**, 713–1018 (2000)
8. Feireisl, E., Gwiazda, P., Świerczewska-Gwiazda, A., Wiedemann, E.: Dissipative measure-valued solutions to the compressible Navier-Stokes system. Calc. Var. Partial Dif. **55**(6), 55–141 (2016)
9. Feireisl, E., Lukáčová-Medvid'ová, M., Mizerová, H.: A finite volume scheme for the Euler system inspired by the two velocities approach. Numer. Math. **144**, 89–132 (2020)
10. Feireisl, E., Lukáčová-Medvid'ová, M., Mizerová, H., She, B.: Convergence of a finite volume scheme for the compressible Navier–Stokes system. ESAIM: M2AN **53**(6), 1957-1979 (2019)
11. Feireisl, E., Lukáčová-Medvid'ová, M., Mizerová, H., She, B.: On the convergence of a finite volume scheme for the compressible Navier–Stokes–Fourier system. IMA J. Numer. Anal. (2020). https://doi.org/10.1093/imanum/draa060
12. Godunov, S.K.: A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. Mat. Sb. (N.S.) **47**(89), 3 271–306 (1959)
13. Guermond, J.L., Popov, B.: Viscous regularization of the Euler equations and entropy principles. SIAM J. Appl. Math. **74**(2), 284–305 (2014)
14. Guermond, J.L., Popov, B.: Invariant domains and first-order continuous finite element approximation for hyperbolic systems. SIAM J. Num. Anal. **54**, 2466–2489 (2016)
15. Liska, R., Wendroff, B.: Comparison of several difference schemes on 1D and 2D test problems for the Euler equations. SIAM J. Sci. Comput. **25**(3), 995–1017 (2003)
16. Lukáčová-Medvid'ová, M., Saibertová, J., Warnecke, G.: Finite Volume Evolution Galerkin Methods for Nonlinear Hyperbolic Systems. J. Comput. Phys. **183**(2), 533–562 (2002)
17. Schulz-Rinne, C.W., Collins, J.P., Glaz, H.M.: Numerical solution of the Riemann problem for two-dimensional gas dynamics. SIAM J. Sci. Comput. **14**(6), 1394–1414 (1993)
18. Shen, H., Wen, C.Y., Zhang, D.L.: A characteristic space-time conservation element and solution element method for conservation laws. J. Comput. Phys. **288**, 101–118 (2015)
19. Shu, C., Osher, S.: Efficient implementation of essentially nonoscillatory shock-capturing schemes. J. Comput. Phys. **77**(2), 439–471 (1988)
20. Tadmor, E.: Entropy stability theory for difference approximations of nonlinear conservation laws and related time dependent problems. Acta Numer. **12**, 451–512 (2003)
21. Tadmor, E.: The numerical viscosity of entropy stable schemes for systems of conservation laws. Math. Comp. **49**(179), 91–103 (1987)
22. Tadmor, E.: Minimum entropy principle in the gas dynamic equations. Appl. Num. Math. **2**, 211–219 (1986)
23. Toro, E.F.: Riemann Solvers and Numerical Methods for Fluid Dynamics. A Practical Introduction, 3rd edn. Springer-Verlag, Berlin (2009)
24. VanderZee, E., Hirani, A.N., Guoy, D., Ramos, E.A.: Well-centered triangulation. SIAM J. Sci. Comput. **31**(6), 4497–4523 (2010)
25. Xu, K., Kim, C., Martinelli, L., Jameson, A.: BGK-based schemes for the simulation of compressible flow. Int. J. Comput. Fluid Dyn. **7**(3), 213–235 (1996)

Recent Advances and Complex Applications of the Compressible Ghost-Fluid Method

Steven Jöns, Christoph Müller, Jonas Zeifang, and Claus-Dieter Munz

Abstract In this paper, improvements to a level-set ghost-fluid scheme in a high order discontinuous Galerkin framework with finite-volume sub-cells are presented. We propose the use of a path-conservative scheme for the level-set transport in both the discontinuous Galerkin and the finite-volume framework. Additionally, improvements regarding the curvature calculation and velocity extrapolation are described. The modified scheme is validated by a comparison of shock-drop and drop-drop interaction simulations from literature.

1 Introduction

Compressible multi-phase flow is of major interest in many scientific and industrial applications. Two major concepts can be distinguished: sharp and diffuse interface methods. Popular sharp interface methods are the volume-of-fluid and the ghost-fluid method. The volume-of-fluid method is widely used for incompressible flow, but was also applied to compressible multi-phase problems, see e.g. [15]. The ghost-fluid method has been introduced by Fedkiw et al. [13] and was improved by many authors, see e.g. Liu et al. [28], [27] and Wang et al. [45]. Merkle and Rhode demonstrated a modified version, where a multi-phase Riemann problem is solved to obtain the ghost states at the interface. The concept was modified to allow approximate two-phase Riemann solvers by Fechter et al. [10–12]. The method was applied within a high order level-set ghost-fluid framework. A discontinuous Galerkin method [17] was used to transport both the bulk phases and the level-set field. Shocks as well as the

S. Jöns, C. Müller and J. Zeifang—These three authors contributed equally to the paper.

S. Jöns · C. Müller · J. Zeifang · C.-D. Munz (✉)
Institute of Aerodynamics and Gas Dynamics at the University of Stuttgart, Pfaffenwaldring 21, 70569 Stuttgart, Germany
e-mail: munz@iag.uni-stuttgart.de

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*, SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2_7

phase boundary between the bulk phases were captured by a finite-volume sub-cell scheme [39]. In [31], the sub-cell shock capturing method was used for the level-set transport as well. Another level-set ghost-fluid approach is based on cut-cell methods, see e.g. [32] or [44]. These approaches use similar methods to handle the level-set and the geometry calculation, however their treatment of the phase boundary is based on cut-cells. While the method of Nourgaliev et al. [32] is non-conservative like the method presented in this paper, Vahab and Miller [44] considered a conservative handling of the phase boundary.

In this paper, we focus on modifications to the interface handling of a compressible sharp interface method in order to simulate merging droplets and bubbles. The numerical framework is based on a discontinuous Galerkin flow solver with finite volume sub-cells for the bulk phases, which are coupled with a level-set ghost-fluid method. The description of the scheme will be kept short, details are described in [11] and [31]. We propose the use of a path-conservative scheme to transport the level-set field. This leads to a modified sub-cell shock capturing based on [8]. We additionally discuss novel modifications of the curvature calculation and the level-set transport, which allow the simulation of phase boundaries with high curvatures as well as topological changes. Afterwards, complex test cases are shown to validate the scheme: two cases with merging drops and two shock-drop interactions, which are compared with results from literature.

2 Governing Equations

The level-set ghost-fluid framework under consideration is a sharp interface method and assumes two distinct pure phases without a mixing zone. We model each of these bulk phases with the compressible Euler equations

$$\frac{\partial \mathbf{Q}}{\partial t} + \nabla \cdot \boldsymbol{\mathcal{F}}(\mathbf{Q}) = 0, \quad \text{with} \quad \mathbf{Q} = \begin{pmatrix} \rho \\ \mathbf{m} \\ E \end{pmatrix} \quad \text{and} \quad \boldsymbol{\mathcal{F}}(\mathbf{Q}) = \begin{pmatrix} \rho\mathbf{u} \\ \rho\mathbf{u}\mathbf{u} + \underline{\mathbf{I}}p \\ (\rho e + p)\mathbf{u} \end{pmatrix}, \quad (1)$$

with density ρ, momentum $\mathbf{m} = \rho\mathbf{u}$ and total energy $E = \rho e$ as conserved quantities. The total energy E is the sum of the internal energy per unit volume $\rho\epsilon$ and the kinetic energy $\frac{1}{2}\rho\mathbf{u} \cdot \mathbf{u}$

$$E = \rho\epsilon + \frac{1}{2}\rho\mathbf{u} \cdot \mathbf{u}, \quad (2)$$

with $\mathbf{u}$ denoting the velocity. An equation of state (EOS) has to be specified to link the pressure and the internal energy per unit mass ϵ:

$$p = p(\rho, \epsilon), \quad \epsilon = \epsilon(\rho, p). \quad (3)$$

With our framework, an arbitrary EOS can be used. Efficiency is secured by using the tabulation technique by Föll et al. [14], whereas an explicit evaluation of algebraic EOS is also possible. In this paper, we fit a stiffened gas EOS, see Saurel et al. [35], for the liquid phase and use the perfect gas law for the gaseous phase. The stiffened gas law is chosen over the Tait EOS although the latter has been found out to model water more precisely, see e.g. [34]. However, the Tait EOS simply links density and pressure. Therefore, it cannot be applied to the full compressible Euler equations directly. Following [13, 29], an additional equation for the internal energy has to be added to use it in this case. With the choice in [13] the Tait EOS can be rewritten to the form of the stiffened gas EOS. This approach was used in e.g. [20, 52] as well. For a further discussion on the use of different EOS for the modeling of water see [5, 20], and e.g. [46, 47] as exemplary applications with different EOS.

The interface between the two phases is tracked by the level-set function Φ, which is transported by a velocity field $\mathbf{s}$ according to

$$\frac{\partial \Phi}{\partial t} + \mathbf{s} \cdot \nabla \Phi = 0. \tag{4}$$

The transport velocity of the level-set function depends on the flow states at the interface $\mathbf{s} = f(\mathbf{Q}_{liq}, \mathbf{Q}_{gas})$. It is initially given on the phase boundary and is then extrapolated into the volume. In our numerical framework (Sect. 3), it is only calculated at the beginning of each time-step to reduce the complexity of the coupling between fluid motion and level-set transport. As a direct consequence, the transport velocity field is constant within each timestep and thus we can rewrite Eq. (4) to

$$\frac{\partial W}{\partial t} + \boldsymbol{\mathcal{B}}(\mathbf{x}) \cdot \nabla W = 0 \quad \text{with} \quad W = \Phi \quad \text{and} \quad \boldsymbol{\mathcal{B}}(\mathbf{x}) = \mathbf{s}. \tag{5}$$

Equation (5) is formulated in the general form of non-conservative hyperbolic equations to introduce the notation for the numerical scheme, which is discussed in Sect. 3. The root of the level-set field marks the phase interface. The level-set function initially fulfills the signed distance property. However, this property is not preserved by the level-set transport (Eq. (5)). As a result, the level-set function needs to be reinitialized. In this work, the method of choice is the solution of a Hamilton-Jacobi equation

$$\frac{\partial \Phi}{\partial t} + \text{sign}(\Phi)\,(|\nabla \Phi| - 1) = 0 \tag{6}$$

as proposed in [40]. There are other approaches to reinitialize the level-set field, see e.g. [37, 43]. A beneficial aspect of level-set methods is that geometrical properties, such as normal vector $\mathbf{n}_{LS}$ and curvature κ_{LS} can be calculated directly from the level-set function by differentiation. According to [6], the level-set normal is calculated by

$$\mathbf{n}_{LS} = \frac{\nabla \Phi}{|\nabla \Phi|}. \tag{7}$$

For the calculation of the curvature the general formulation given in [6]

$$\begin{aligned} \kappa_{LS} = & \frac{\Phi_x^2 \Phi_{yy} - 2\Phi_x \Phi_y \Phi_{xy} + \Phi_y^2 \Phi_{xx}}{|\nabla \Phi|^3} + \\ & \frac{\Phi_x^2 \Phi_{zz} - 2\Phi_x \Phi_z \Phi_{xz} + \Phi_z^2 \Phi_{xx}}{|\nabla \Phi|^3} + \\ & \frac{\Phi_y^2 \Phi_{zz} - 2\Phi_y \Phi_z \Phi_{yz} + \Phi_z^2 \Phi_{yy}}{|\nabla \Phi|^3}. \end{aligned} \tag{8}$$

is preferred over the simpler formulation

$$\kappa_{LS} = \nabla \cdot \mathbf{n}_{LS}. \tag{9}$$

We found that the general formulation is beneficial to obtain stable simulations of merging droplets. A possible reason is the underresolution of the level-set field in these situations. More sophisticated algorithms for the normal and curvature calculation based on curve parametrizations are discussed e.g. in [26], but are not considered in this work due to their increased computational cost. In addition to the geometrical properties, the velocity of the level-set field $\mathbf{s}$ has to be determined as well. It is only calculated on the phase boundary and has to be extrapolated into the volume. This is typically done in a two-step procedure: First, the data is set in the neighborhood of the phase boundary. Afterwards, this initial field is extrapolated by solving the Hamilton-Jacobi equations

$$\frac{\partial s^i}{\partial \tau} + \text{sign}(\Phi)\mathbf{n}_{LS} \cdot \nabla s^i = 0 \quad i = 1, \ldots, d, \tag{10}$$

for the direction-wise components s^i of the d-dimensional velocity field $\mathbf{s}$ following [1]. This is discussed in more detail in Sect. 3.2.3.

Both the reintialization and the velocity extrapolation are only performed in a narrow, radial band around the level-set zero. Outside the narrow band the level-set function is set to the bands fixed radius and the velocity field is set to zero.

3 Numerical Method

In this section, the general numerical framework is described briefly. First, the building blocks for the method are described in Sect. 3.1. Afterwards, they are assembled in Sec. 3.2 to form the high order framework for the sharp interface simulations.

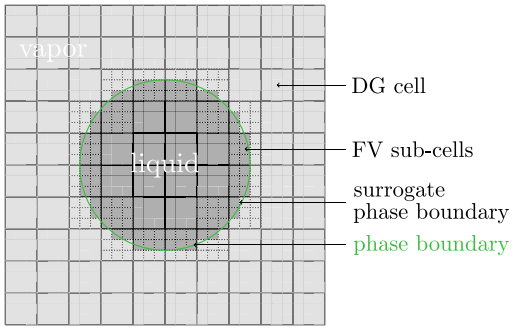


Fig. 1 Domain decomposition into liquid (dark gray) and vapor (light gray) region, using the zero position of a level-set function (green). Instead of the DG method, a FV sub-cell scheme is used in a narrow band around the resulting surrogate phase boundary.

We start by introducing basic notation. The domain $\overline{\Omega}$ is divided into a liquid region Ω^l and a vapor region Ω^v with the phase interface Γ. It holds

$$\overline{\Omega} = \overline{\Omega^l} \cup \overline{\Omega^v}, \ \Omega^l \cap \Omega^v = \emptyset, \ \text{and} \ \overline{\Omega^l} \cap \overline{\Omega^v} = \Gamma.$$

Furthermore, $\overline{\Omega}$ is discretized with hexahedral elements such that

$$\overline{\Omega} = \bigcup_e \overline{\Omega_e}, \ \text{and} \ \Omega_i \cap \Omega_j = \emptyset, \quad \forall i \neq j.$$

The numerical framework used in this work is based on [11] and [31]. Liquid and vapor region are both discretized with the discontinuous Galerkin method. At the phase boundary, which is defined by the zero position of a level-set function, a finite volume sub-cell scheme is applied to ensure a better representation of the surrogate phase boundary. This surrogate surface discretely represents the phase boundary and is aligned with the sub-cell interfaces. For an overview on the domain decomposition see Fig. 1.

3.1 Building Blocks for the Level-Set Ghost-Fluid Method

3.1.1 The DGSEM with Finite-Volume Sub-Cells

In this subsection, we discuss the Discontinuous Galerkin Spectral Element Method (DGSEM) [23] with finite-volume sub-cells [18, 33, 39] for hyperbolic conservation laws. We extend the DGSEM with finite-volume sub-cells to the case of hyperbolic equations with non-conservative products, following [8]. Therefore, the framework of path-conservative schemes is used [4]. In the DGSEM, the approximate solution of both the bulk phases and the level-set is described by piecewise polynomials $\mathbf{Q}_h$ and W_h of degree N, respectively. Within each element Ω_e the solutions are represented by a tensor product of nodal one-dimensional Lagrange basis functions. The basis

functions are chosen to be identical to the test functions l in the weak formulations. The weak formulations for Eqs. (1) and (5) read

$$\frac{\partial}{\partial t}\int\limits_{\Omega_e} \mathbf{Q}_h l d\mathbf{x} + \oint\limits_{\partial\Omega_e} \boldsymbol{\mathcal{F}}(\mathbf{Q}_h)\cdot\mathbf{n} l ds - \int\limits_{\Omega_e} \boldsymbol{\mathcal{F}}(\mathbf{Q}_h)\cdot\nabla l d\mathbf{x} = 0, \tag{11}$$

$$\frac{\partial}{\partial t}\int\limits_{\Omega_e} W_h l d\mathbf{x} + \oint\limits_{\partial\Omega_e} \boldsymbol{\mathcal{B}}(\mathbf{x})\cdot\nabla W_h l ds + \int\limits_{\Omega_e} \boldsymbol{\mathcal{B}}(\mathbf{x})\cdot\nabla W_h l d\mathbf{x} = 0, \tag{12}$$

with the outward pointing normal vector $\mathbf{n}$. In the Euler equations, the neighboring elements are coupled by a numerical flux function $\boldsymbol{\mathcal{F}}^*(\mathbf{Q}_h^-, \mathbf{Q}_h^+)\cdot\mathbf{n} \approx \boldsymbol{\mathcal{F}}(\mathbf{Q}_h)\cdot\mathbf{n}$. We use the HLLC [42] and the HLLE [9] Riemann solver. For the level-set transport equation, the path-conservative jump term $\boldsymbol{\mathcal{D}}^*(W_h^-, W_h^+)\cdot\mathbf{n} \approx \boldsymbol{\mathcal{B}}(\mathbf{x})\cdot\nabla W_h$ has to be approximated. We use the path-conservative Rusanov Riemann solver [8]

$$\boldsymbol{\mathcal{D}}^*(W_h^-, W_h^+)\cdot\mathbf{n} = \frac{1}{2}\left(\widetilde{\boldsymbol{\mathcal{B}}}\cdot\mathbf{n} - s_{max}\underline{\mathbf{I}}\right)\left(W_h^+ - W_h^-\right) \tag{13}$$

with the maximal signal speed

$$s_{max} = \max\left(|\mathbf{s}^+\cdot\mathbf{n}|, |\mathbf{s}^-\cdot\mathbf{n}|\right). \tag{14}$$

The superscript $(\cdot)^-$ identifies the value inside the current cell and the superscript $(\cdot)^+$ identifies the value outside the current cell. To approximate the Roe type matrix $\widetilde{\boldsymbol{\mathcal{B}}}$ we substitute the spatial dependency of $\boldsymbol{\mathcal{B}}$ on $\mathbf{x}$ with a dependency on W_h. In general, this is not valid as the advection field is a function of space. However, the level-set variable Φ carries the signed-distance property. Hence, in a 1D Riemann problem it is possible to transform the spatial dependency to a dependency on the level-set field. This enables to approximate $\widetilde{\boldsymbol{\mathcal{B}}}$ by integrating along a linear path $\Psi(W_h^-, W_h^+, b)$, with $b \in [0, 1]$ between W_h^- and W_h^+ as

$$\widetilde{\boldsymbol{\mathcal{B}}}\cdot\mathbf{n} \approx \int\limits_0^1 \boldsymbol{\mathcal{B}}(\Psi(W_h^-, W_h^+, b))\cdot\mathbf{n} db, \quad \Psi(W_h^-, W_h^+, b) = W_h^- + b(W_h^+ - W_h^-). \tag{15}$$

We evaluate the path numerically with the trapezoidal rule and obtain

$$\widetilde{\boldsymbol{\mathcal{B}}}\cdot\mathbf{n} \approx \frac{\boldsymbol{\mathcal{B}}(W_h^-) + \boldsymbol{\mathcal{B}}(W_h^+)}{2}\cdot\mathbf{n}. \tag{16}$$

The volume terms in Eq. (11) and Eq. (12) can be calculated directly from $\mathbf{Q}_h$ and W_h. The derivatives of the test function and the solution can be calculated by derivating the respective polynomial.

The main idea of the DGSEM is to choose the same $N+1$ Legendre-Gauss points for both the numerical integration and the interpolation of the solution. This reduces the number of operations per degree of freedom and increases the efficiency. In addition, the multi-dimensional operator simplifies to a subsequent application of the one-dimensional operator. Details on the implementation can be found in [2, 24] and [17].

The finite-volume formulations of Eqs. (1) and (5) are a special case of the weak formulations Eqs. (11) and (12). If both the solution and the testfunction are chosen out of the space of piecewise constant polynomials, e.g. $l = 1$, Eqs. (11) and (12) simplify to the finite-volume methods

$$\frac{\partial}{\partial t} \int_{\Omega_e} \mathbf{Q}_h d\mathbf{x} + \oint_{\partial\Omega_e} \boldsymbol{\mathcal{F}}^*(\mathbf{Q}_h^-, \mathbf{Q}_h^+) \cdot \mathbf{n} ds = 0, \tag{17}$$

$$\frac{\partial}{\partial t} \int_{\Omega_e} W_h d\mathbf{x} + \oint_{\partial\Omega_e} \boldsymbol{\mathcal{D}}^*(W_h^-, W_h^+) \cdot \mathbf{n} ds = 0. \tag{18}$$

We combine the DGSEM and FV approach to capture discontinuities in the high order DGSEM solution, which would otherwise lead to oscillations. In the Euler equations, shocks and the phase boundary have to be captured. For the level-set equation, the edge of the narrow band requires stabilization. In addition, areas with a high curvature with respect to the grid resolution are troublesome. This can be resolved by either a grid refinement or the sub-cell scheme. Additional problems occur if level-set contours merge, e.g. merging drops. In this case, the process has to be captured by a low order scheme like the sub-cell approach. For the sub-cell method we formulate an a priori limiter following [39] in contrast to the a posteriori limiter in [8]. The biggest advantage is a reduction in computational cost, since only one operator is evaluated in each cell. A disadvantage is that there is no guarantee of a stable solution. We combine multiple approaches to identify trouble cells of the DGSEM method: First, a modal indicator following [18, 33] is used to detect strong gradients in the solution of the Euler equations and the edge of the narrow band in the solution of the level-set function. Details on modifications and the implementation can be found in [38]. Second, the position of the level-set zero is used to capture the phase boundary in the Euler equations. Finally, we detect zones in which two phase boundaries meet, e.g. the merging of two droplets. Details are discussed in Sect. 3.2.4. After the troubled cells are identified, the polynomial solution is switched to a finite-volume representation. If the cells are no longer problematic they are switched back. The switch upholds

$$\int_{\Omega_e} \mathbf{U} d\mathbf{x} \equiv \int_{\Omega_e} \mathbf{U}_{DG} d\mathbf{x} = \int_{\Omega_e} \mathbf{U}_{FV} d\mathbf{x} \quad \mathbf{U} = \mathbf{Q}, W \tag{19}$$

and hence is conservative. It can be formulated as a matrix vector multiplication with an integration matrix $\mathcal{V}$

$$\mathcal{V}\mathbf{U}_{DG} = \mathbf{U}_{FV} \quad \mathbf{U} = \mathbf{Q}, W. \tag{20}$$

The polynomial representation has $(N+1)^d$ degrees of freedom, with d denoting the number of dimensions. For the finite-volume method, we choose to use $(N+1)^d$ equidistantly distributed finite-volume sub-cells. This choice allows the use of the same data structure and hence an easy implementation. For the Euler equation, the finite-volume scheme is extended to a second order TVD scheme. The coupling between the DGSEM and the sub-cells occurs via the surface terms on the element boundaries. The fluxes and jump terms are evaluated in the finite-volume discretization and then projected to the polynomial discretization for the DG elements. The scheme handles discontinuities well, since it switches to the finite-volume representation if necessary. Its use of finite-volume sub-cells intrinsically leads to a grid refinement, which prevents a strong loss of accuracy.

We want to highlight some advantages of the novel approach for the level-set transport compared with the previously discussed approach in [12, 31]. There, Eq. (5) is used in a divergence form with a source term

$$\frac{\partial \Phi}{\partial t} + \nabla \cdot (\mathbf{s}\Phi) = \Phi \nabla \cdot \mathbf{s}. \tag{21}$$

In the incompressible case the right hand side of Eq. (21) is zero since the velocity field is divergence free ($\nabla \cdot \mathbf{s} = 0$), see e.g. [16, 30]. However, in the compressible case this term needs to be discretized.

The novel path-conservative scheme has two advantages: First, the DGSEM and finite-volume sub-cell scheme are derived from the same weak formulation and solve the same equation on the discrete level. In this sense, they are consistent. If Eq. (21) is solved instead, $\nabla \cdot \mathbf{s} \neq 0$ still holds discretely for the discontinuous Galerkin method. However, in the finite-volume scheme $\mathbf{s}$ is discretized with constant polynomials and thus $\nabla \cdot \mathbf{s} = 0$. As a result both schemes solve different equations although they are formally derived from the same weak formulation. Secondly, the time-step is only limited by the eigenvalues of the hyperbolic transport, which is the process of interest. If Eq. (21) is solved, the source term might be stiff. In this case the eigenvalues of the source are larger than those of the hyperbolic transport. This leads to smaller time-steps and thus higher costs of the numerical simulations.

3.1.2 Time Discretization

For the temporal discretization of the level-set transport and the Euler equations we either use a fully explicit 4^{th} order Runge-Kutta (RK) scheme from [3] or an implicit-explicit 4^{th} order Runge-Kutta scheme by Kennedy and Carpenter [21]. The goal of implicit-explicit time discretization is to overcome the severe time step restriction

of explicit schemes at low Mach numbers. Hence, we treat the Euler equations with an implicit scheme and the level-set transport with an explicit scheme. For solving the arising non-linear equation system of the implicit part we rely on a matrix-free Newton-GMRES approach [22] as it is applied to the DGSEM formulation in [50, 51] and extend it to the mixed DG-FV ghost-fluid formulation. More details about this time discretization will be presented in a follow-up publication.

3.2 The Level-Set Ghost-Fluid Method

The numerical methods described above are the building blocks of the present level-set ghost-fluid framework. In the following, we describe the necessary steps to assemble the framework.

3.2.1 Algorithmic Details of the Level-Set Ghost-Fluid Method

We follow the approach in [11] for the development of our method. It consists of the repetition of the following steps:

1. The RK-DGSEM/RK-FV solver is used to advance the level-set field and the Euler equations for the pure phases in time.
2. The level-set function is reinitialized.
3. Depending on the level-set root, the domain Ω is decomposed into Ω^l and Ω^v along the boundaries of the finite-volume sub-cells using already existing physical states and ghost states. This creates a surrogate phase boundary Γ^s.
4. The DG-FV distribution of both the Euler equations and the level-set is updated based on modal smoothness indicators and geometrical information of the level-set function.
5. The normal vector at the phase interface and the curvature are calculated.
6. The boundary conditions at the surrogate phase boundary and its velocity are calculated with a two-phase Riemann solver, the so called HLLP Riemann solver [11, 36]. It models surface tension with a jump term across the phase boundary and provides both fluxes for each phase as well as the velocity of the phase boundary.
7. The interface velocity is then extrapolated into the volume to obtain a velocity field for the level-set transport.

Before the initial time-step is executed, the above mentioned procedure has to be done once without step 1. Additionally, step 6 is applied in each Runge-Kutta stage. The presented level-set ghost-fluid algorithm does not guarantee conservation due to two reasons. First, the fluxes at the surrogate phase boundary may be distinct to ensure a stable two-fluid simulation. Secondly, the state of cells which change from the liquid into the vapor domain and vice versa are replaced with their respective ghost state in step 3. For more details about the method the reader is referred to [31].

3.2.2 Calculation of Derivatives: The Level-Set Normal Vector and Curvature

The normals and the curvature are calculated by first transforming the level-set solution to the finite-volume sub-cell representation via Eq. (20) and secondly calculating the derivatives of the level-set function with a 5th order WENO stencil as proposed in [10]. The same operator is applied again to the components of $\nabla\Phi$ to obtain the second order derivatives of the level-set field. With these gradients we then evaluate Eq. (7) and Eq. (8) to calculate normal vectors and curvature of the level-set function. We additionally limit the curvature by an upper bound that depends on the grid resolution, which we characterize by $r^{min} = \min(V_{FV})^{1/d}$. Thereby, V_{FV} is the volume of a finite-volume sub-cell. With this we can define

$$|\kappa_{LS}|^{max} = \frac{d-1}{2r^{min}} = \frac{d-1}{2\min(V_{FV})^{1/d}}, \tag{22}$$

as the maximum absolute value of the curvature, which can be resolved by the grid assuming a safety factor of 0.5.

3.2.3 Solution of Hamilton-Jacobi Equations: Reinitialization and Velocity Extrapolation

Two sets of Hamilton-Jacobi equations need to be solved: the reinitialization equation Eq. (6) and the equations for the velocity extrapolation Eq. (10). Each set of equations is solved with a 5th order WENO scheme [19] in combination with a third order order low storage Runge-Kutta method with three stages [48]. For the velocity extrapolation an additional step is necessary. The solution of the two-phase Riemann problem gives a transport velocity on the cell edges that form the surrogate phase boundary. This velocity can be directly copied to the neighboring finite-volume sub-cells. If a sub-cell is involved in more than one two-phase Riemann problem, we average the velocities. In the direct neighbors of the surrogate phase interface, the velocity is fixed. In all other cells within the narrow band, we solve Eq. (10) to obtain a smooth velocity field.

3.2.4 Specific Modifications of the Algorithm for Simulating Merging Droplets

If topological changes are simulated, e.g. merging droplets, special attention has to be payed in regions where those topological changes take place. In the following we detail the two necessary modifications.

In a first step, we have to ensure that we capture topological changes with the finite-volume sub-cell framework, since they are associated with discontinuities in the level-set field. We use the topological information that is available through the

sign of the level-set function. Therefore, we switch the level-set solution to the finite-volume sub-cell discretization in each element. Afterwards, we evaluate the level-set sign line-wise in all spatial directions. If the sign changes more than once, the DG element contains a topological change which has to be captured with the sub-cell approach. In a second step, we have to identify the specific sub-cells that are involved in the topological change. We check the number of two-phase Riemann problems a sub-cell is involved in. If a sub-cell is affected by more than one two-phase Riemann problem per direction it is identified as a potential *merge cell* ($\mathcal{I}_{\text{topo}} = 1$). In those cells, the advection velocity cannot be determined by averaging, cf. Sect. 3.2.3. The use of the advection velocity in the *merge cells* is avoided by introducing a specific form of the path-conservative jump term. Summarizing Eqs. (13)-(16), the path conservative jump term $\mathcal{D}^*(W_h^-, W_h^+)$ for the transport of the level-set field Eq. (5) is

$$\mathcal{D}^*(\Phi^-, \Phi^+) = \frac{1}{4}((s^+ + s^-) - 2\ \max\left(|s^+|, |s^-|\right))(\Phi^+ - \Phi^-), \tag{23}$$

where the velocities s^+, s^- are selected via

$$\begin{aligned} s^- &= \mathbf{s}^- \cdot \mathbf{n},\ s^+ = \mathbf{s}^+ \cdot \mathbf{n} \quad \text{if} \quad \mathcal{I}^-_{\text{topo}} = \mathcal{I}^+_{\text{topo}}, \\ s^- &= \mathbf{s}^- \cdot \mathbf{n},\ s^+ = \mathbf{s}^- \cdot \mathbf{n} \quad \text{if} \quad \mathcal{I}^-_{\text{topo}} = 0,\ \mathcal{I}^+_{\text{topo}} = 1, \\ s^- &= \mathbf{s}^+ \cdot \mathbf{n},\ s^+ = \mathbf{s}^+ \cdot \mathbf{n} \quad \text{if} \quad \mathcal{I}^-_{\text{topo}} = 1,\ \mathcal{I}^+_{\text{topo}} = 0. \end{aligned} \tag{24}$$

Hence, the transport velocity is chosen according to the $\mathcal{I}_{\text{topo}}$ identifier. This procedure ensures that the velocity for the level-set transport is taken only from cells that are not involved in a topological change.

4 Numerical Results

In this section we apply the numerical framework to test problems in the low and high Mach number regime. First, we look at two droplets in a gas with a linear velocity profile to evaluate the accuracy of the curvature calculation. Secondly, a simulation of two merging droplets with a Weber number of $\text{We} = 2.2$ is performed. Afterwards, two shock-drop interactions with a shock Mach number of $\text{M}_s = 1.47$ and Weber numbers of $\text{We} = 7339$ and $\text{We} = 12$ are simulated and compared to results from literature.

4.1 Drop Collision in Linear Velocity Profile

The first testcase is an adaption from [25] in which a drop collision in a shear layer has been described. Since we only consider inviscid fluids in this paper, we have slightly

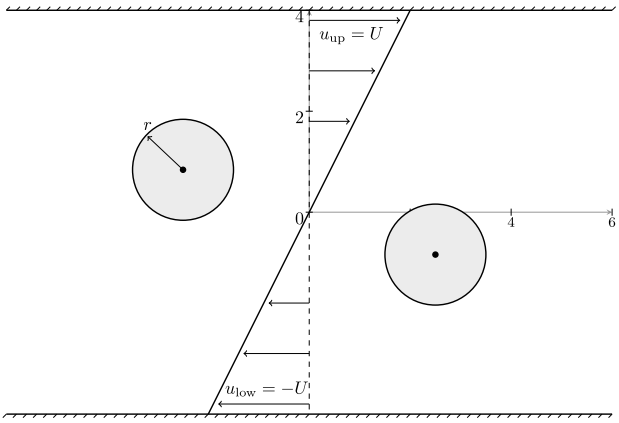


Fig. 2 Computational setup for a drop collision in a gas with a linear velocity profile. The upper and lower boundaries are slip walls whereas the left and right boundaries are periodic. The domain is discretized with 90×60 4^{th} order elements coupled with a HLLC Riemann solver.

Table 1 Initial conditions and material parameters for the drop collision in a linear velocity profile.

	p_0	ρ_0	γ	p_∞	σ
Gas phase	71.43	1.0	1.4	0	–
Liquid phase	81.43	1.0	7.15	3307	10

altered the test case into a collision of two drops emerged in a gas with a linear velocity profile. It allows to benchmark the curvature calculation during the merging process. It is a difficult test, since the droplets have only a small relative velocity in normal direction. The setup is visualized in Fig. 2 with the following parameters: radius $r = 1$, position of the left drop $(-2.5, 0.84)$, position of the right drop $(2.5, -0.84)$ and maximum velocity $U = 1$. The initial conditions and material parameters are given in Table 1. For the time discretization an explicit 4^{th} order Runge-Kutta scheme with CFL $= 0.3$ is used. If the static capillary time step restriction is calculated as in [7], the time-step is limited by the wave propagation of the acoustic waves. The ratio of capillary to accoustic time-step is $\Delta t^{\text{capillary}} / \Delta t^{\text{acoustic}} \approx 24$.

In Fig. 3 the temporal evolution of the phase boundary during a collision is visualized. Due to the linear velocity profile, the drops are deformed as they approach each other. The distance between the two drops shrinks until it can no longer be resolved by the grid. At this instance the drops merge. We observe two merges at two different positions that happen almost at the same time. They enclose a vapor bubble, which vanishes quickly due to its underresolution. Afterwards, a wave moves along the drop surface and changes the droplets shape towards a spherical form. The same qualitative behavior is observed in [25]. We conclude that our numerical framework is suitable to simulate merge phenomena.

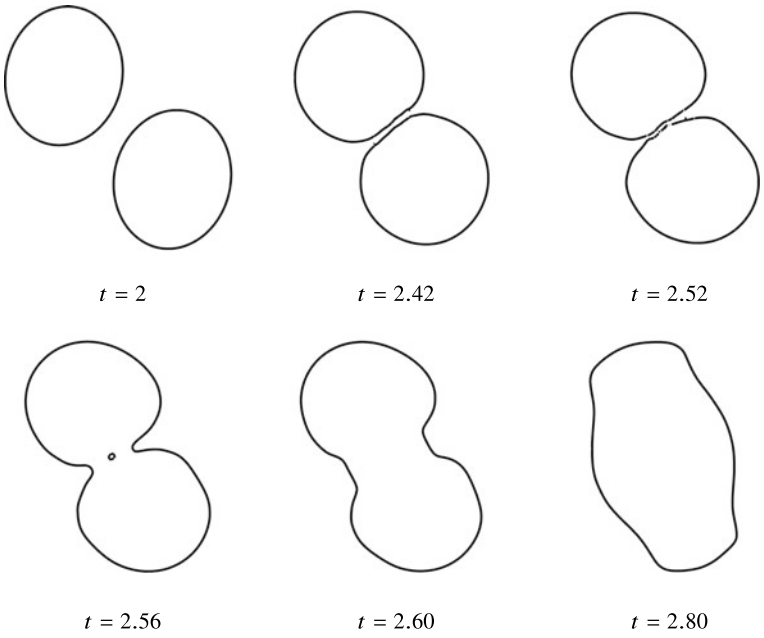


Fig. 3 Temporal evolution of the phase boundary for the drop collision in linear velocity profile. Due to inertial forces the drops are deformed as they approach each other. Then they merge and form a single drop.

Table 2 Initial conditions and material parameters for collision of liquid ethanol droplets in air.

	p_0[bar]	ρ_0[kg m^{-3}]	γ[−]	p_∞[bar]	σ[kg s^{-2}]
Air	1.0	1.226	1.4	0	–
Ethanol$_l$	1.11375	791	1.208	8466.14	2.275

4.2 Droplet Collisions with $\mathbf{We}_c = 2.2$

In a next step, we apply our framework to a binary droplet collision. Inspired by [41], we simulate ethanol droplets in air with a collision Weber number of $\mathrm{We}_c = 2.2$, being defined as

$$\mathrm{We}_c = \frac{\rho_l U_c^2 d}{\sigma}, \tag{25}$$

with the liquid density ρ_l, the droplet diameter d and the relative velocity of the droplet U_c. The initial conditions and material parameters are summarized in Table 2.

Both drops have a radius of $r = 0.2$ mm. Initially, they are seperated by a distance of $2.5r$. The droplets are initialized with a vertical velocity of $v^{(1)} = 2.0\frac{\mathrm{m}}{\mathrm{s}}$ and

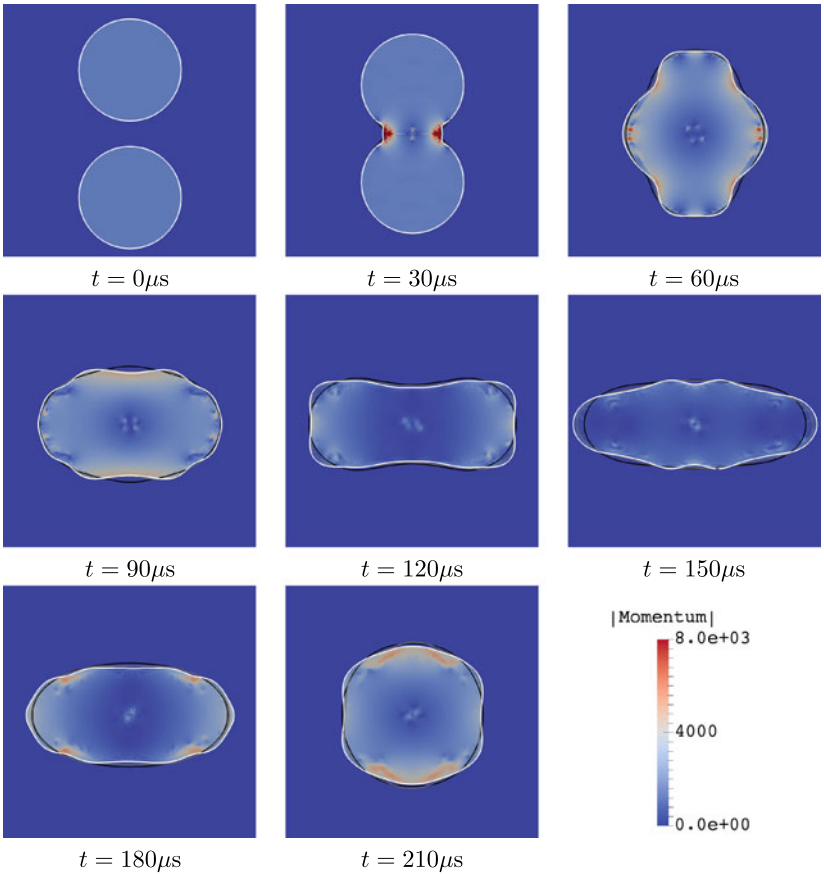


Fig. 4 Temporal evolution of the phase boundary (white line) and the absolute value of the momentum for the drop collision with $We_c = 2.2$, choosing 128^2 elements with $N = 4$ for the spatial resolution. The gray and the black line indicate the phase boundary of the simulations with 64^2 and 32^2 elements, respectively.

$v^{(2)} = -2.0\frac{\mathrm{m}}{\mathrm{s}}$ to obtain a collision Weber number of $We_c = 2.2$. Following [41], we expect coalescence of the two droplets and an oscillation of the remaining single drop. Diverging from the setup in [41] we do not use radial coordinates and neglect viscous effects. The domain $\Omega = [-0.75mm, 0.75mm]^2$ is discretized with three different resolutions with 32^2, 64^2 and 128^2 elements. A polynomial degree of $N = 4$ and the HLLC Riemann solver is used. A 4^{th} order implicit-explicit scheme, see Sect. 3.1.2, is used for the time discretization. For this setup, we achieve a speed-up of approximately 4 compared to the fully explicit scheme. Due to the low Mach number, we can choose a time-step that is approximately 25 times larger than for the explicit scheme with a CFL number of CFL = 0.8. Still, the acoustic waves are the

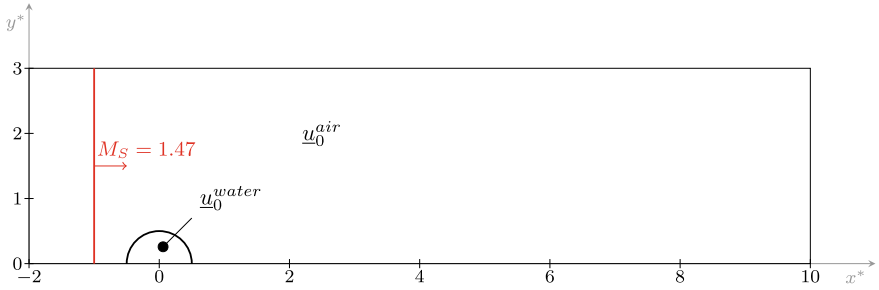


Fig. 5 Computational setup for the shock droplet interaction assuming symmetric conditions. The shock is depicted in red.

fastest characteristics as the ratio $\Delta t^{\text{capillary}}_{\text{explicit}} / \Delta t^{\text{acoustic}}_{\text{explicit}}$ ranges from ≈ 75 to ≈ 150 for the three different spatial resolutions.

In Fig. 4, the temporal evolution of the absolute value of the momentum for the discretization with 128^2 elements is visualized. Additionally, the phase boundary for the simulations with 32^2, 64^2 and 128^2 elements is indicated with a black, a gray and a white line, respectively. We observe the expected behavior qualitatively. A quantitative comparison with [41] is not possible due to the neglect of viscous and three-dimensional effects. A particular consequence of the inviscid flow model is the occurence of further deformations of the phase interface under grid refinement. The only stabilizing mechanism is the numerical viscosity, which decreases as the grid resolution increases. A convergent behavior of the observed phenomena should occur if viscous effects are considered.

4.3 Shock-Droplet Interaction

In the following, we simulate 2D shock-droplet interactions at two different Weber numbers, $\text{We} = 7339$ and $\text{We} = 12$. The numerical setup is taken from Winter et al. [49] and is visualized in Fig. 5. However, we neglect viscous effects. We initialize a water droplet at rest surrounded by air at $x = 0$. A right moving shock wave with a Mach number of $\text{M}_s = 1.47$ is positioned at $x = -D_0$. The initial droplet diameter is chosen as $D_0 = 1m$. The lower domain boundary is set as a symmetry plane. On the left, Dirichlet boundary conditions impose the initial conditions onto the boundary. The remaining boundaries are treated as supersonic outflows. Since both M_s and We are higher than in the test cases considered in Sects. 4.1 and 4.2, the capillary time-step restriction is not considered here. The domain $\Omega = [-2D_0, 10D_0] \times [0, 3D_0]$ is discretized with 512×256 DG elements. For this testcase we use the HLLE Riemann solver, explicit 4^{th} order Runge-Kutta time integration and a polynomial degree of $N = 3$. Initial conditions and material parameters for both considered cases are given in Table 3. The droplet is initialized in mechanical equilibrium with the surrounding

Table 3 Initial conditions and material parameters for the shock-droplet interaction.

	p_0[bar]	ρ_0[kg m^{-3}]	γ[−]	p_∞[bar]	σ[Nm^{-1}]
Air	1.01325	1.204	1.4	0	–
Water (SIE)	1.01355	1000	6.12	3430	15.1571
Water (RTP)	1.19865	1000	6.12	3430	9269.85

air. The pressure difference between droplet and air is given by the Young-Laplace Law. For the comparison with the literature, the non-dimensional time t^* is defined as

$$t^* = \frac{t}{\frac{D_0}{u_s}\sqrt{\frac{\rho_L}{\rho_s}}}, \tag{26}$$

with u_s denoting the post shock velocity and ρ_s the post shock density.

As in Winter et al. [49], we considered two breakup regimes: the shear induced entrainment (SIE) regime and the Rayleigh-Taylor piercing (RTP) regime. For the SIE case, We = 7339, we show the results at the time instances $t^* = 0.25$, $t^* = 0.75$ and $t^* = 1.5$ in Fig. 6. After the shock impinges on the droplet, the surrounding flow deforms the droplet's surface and a complex vortex system is generated in the wake. The two disconnected ligaments at $t^* = 1.5$ stem from the fact that part of the interface has already left the domain. Comparing with the results of Winter et al. [49], the deformations are very similar in the early stages of the simulation. At later stages, differences become more and more apparent. These stem from the neglect of viscous effects in the presented results. The inclusion of viscosity at the interface by Winter et al. [49] produces a smoother droplet surface and a postponed breakup. This can be observed by comparing the time instance $t^* = 1.5$ from Fig. 6 with their results. Nevertheless, this testcase displays the capability of the proposed framework to capture strongly deformed interfaces. An inclusion of viscous effects will be considered in future work.

Next, we consider the RTP case, We = 12. Here, viscous effects are negligible. Results for the non-dimensional time instances $t^* = 0.25$, $t^* = 0.75$ and $t^* = 1.5$ can be seen in Fig. 7. Similar to the SIE case, the droplet deforms after the shock impingement. However, due to the larger surface tension forces, the droplet maintains a more compact form in contrast to the case with a higher We number. Comparing with the results shown in [49], both simulations show a good agreement in the predicted droplet shape. These results demonstrate that our method allows to simulate complex high Mach number settings.

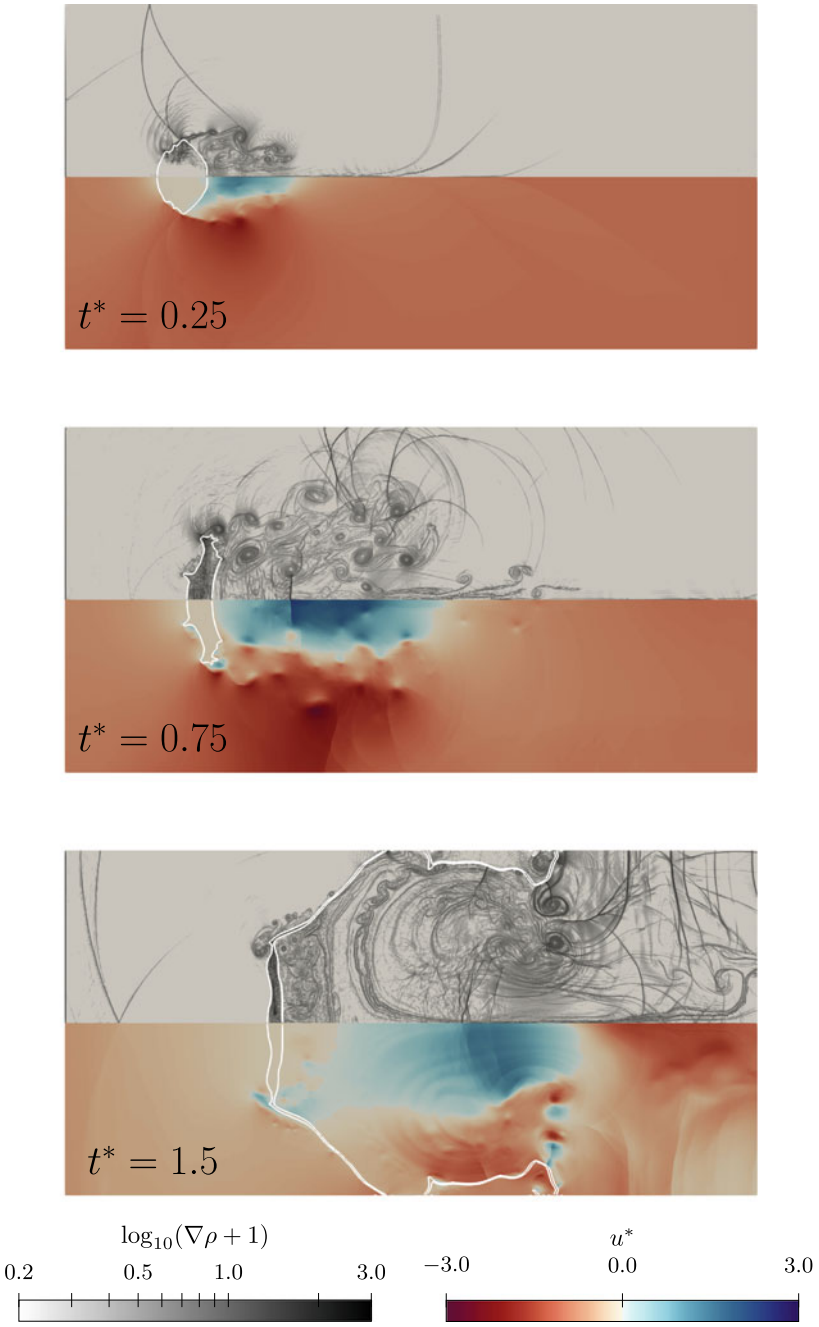


Fig. 6 Numerical Schlieren image(top) and non-dimensional streamwise velocity(bottom) $u^* = u/u_s$ for the SIE case at different time instances. The phase interface is depicted in white.

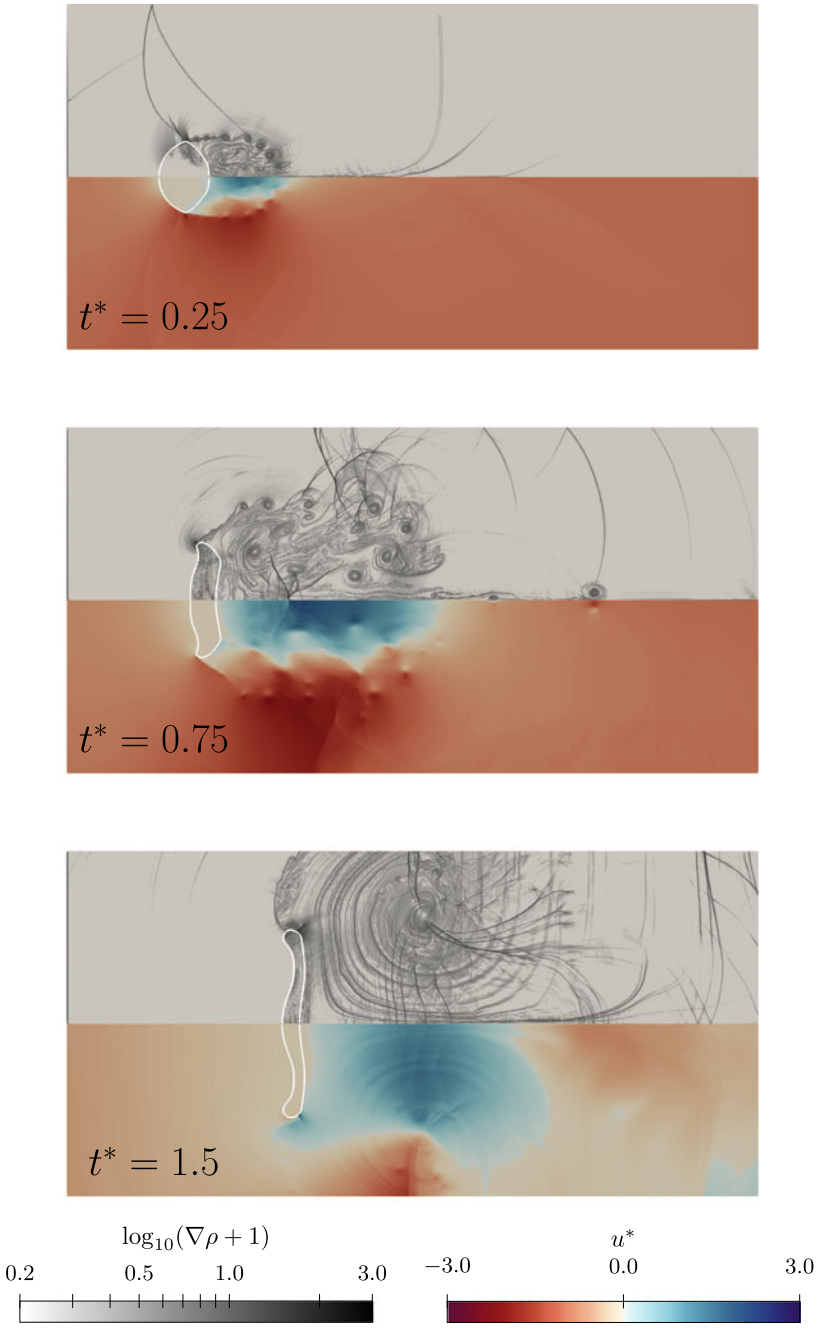


Fig. 7 Numerical Schlieren image(top) and non-dimensional streamwise velocity(bottom) $u^* = u/u_s$ for the RTP case at different time instances. The phase interface is depicted in white.

5 Conclusion

In this paper we provided an overview over a level-set ghost-fluid framework for sharp interface multi-phase flow simulations based on the work in [10] and [31]. We discussed an improved finite-volume sub-cell scheme for the level-set equation based on path-conservative schemes. In addition, two changes in the curvature calculation were presented. At first, the general second derivative of the level-set was used. Secondly, we introduced an upper bound of the curvature value, depending on the size of the grid elements, limiting the curvature by the grid resolution. At last, we discussed a necessary modification of the level-set transport to capture merging phenomena. We avoided the use of the transport velocity in the *merge cells* since it cannot be defined properly.

We showed that these modifications allow the simulation of merging drops with surface tension in two settings: drops in a gas with a linear velocity profile and colliding drops. Afterwards we showed that complex shock-drop interactions are also well within the capabilities of the framework. The modifications allowed the resolution of very fine two-phase structures with respect to the grid size and ensured a stable simulation. We currently work on an extension of the framework to viscous flows. In addition, more complex interactions of bubbles and droplets and drop-wall interactions will be addressed. Detailed investigations on the implicit-explicit framework are currently underway.

Acknowledgements C. Müller and J. Zeifang were supported by the German Research Foundation (DFG) through the project GRK 2160/1 "Droplet Interaction Technologies". S. Jöns was supported by the DFG through the project SFB-TRR 75, Project number 84292822 - "Droplet Dynamics Under Extreme Ambient Conditions" and C.-D. Munz by the DFG under Germany's Excellence Strategy - EXC 2075 - 390740016. The simulations were performed on the national supercomputer Cray XC40 (Hazel Hen) at the High Performance Computing Center Stuttgart (HLRS) under the grant numbers *hpcmphas/44084*.

References

1. Aslam, T.D.: A partial differential equation approach to multidimensional extrapolation. J. Comput. Phys. **193**(1), 349–355 (2004). https://doi.org/10.1016/j.jcp.2003.08.001
2. Beck, A.D., Bolemann, T., Flad, D., Frank, H., Gassner, G.J., Hindenlang, F., Munz, C.D.: High-order discontinuous Galerkin spectral element methods for transitional and turbulent flow simulations. Int. J. Numer. Methods Fluids **76**(8), 522–548 (2014)
3. Carpenter, M., Kennedy, C.: Fourth-order 2N-storage Runge-Kutta schemes. Technical report, NASA Langley Research Center (1994)
4. Castro, M., Gallardo, J., Parés, C.: High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative products. Applications to shallow-water systems. Mat. Comput. **75**(255), 1103–1134 (2006)
5. Chandran, J., Salih, A.: A modified equation of state for water for a wide range of pressure and the concept of water shock tube. Fluid Phase Equilib. **483**, 182–188 (2019)

6. du Chéné, A., Min, C., Gibou, F.: Second-order accurate computation of curvatures in a level set framework using novel high-order reinitialization schemes. J. Sci. Comput. **35**(2–3), 114–131 (2007). https://doi.org/10.1007/s10915-007-9177-1
7. Denner, F., van Wachem, B.G.: Numerical time-step restrictions as a result of capillary waves. J. Comput. Phys. **285**, 24–40 (2015)
8. Dumbser, M., Loubère, R.: A simple robust and accurate a posteriori sub-cell finite volume limiter for the discontinuous Galerkin method on unstructured meshes. J. Comput. Phys. **319**, 163–199 (2016). https://doi.org/10.1016/j.jcp.2016.05.002
9. Einfeldt, B.: On Godunov-type methods for gas dynamics. SIAM J. Numer. Anal. **25**(2), 294–318 (1988). https://doi.org/10.1137/0725021
10. Fechter, S.: Compressible multi-phase simulation at extreme conditions using a discontinuous galerkin scheme (2015). https://doi.org/10.18419/opus-3982
11. Fechter, S., Jaegle, F., Schleper, V.: Exact and approximate Riemann solvers at phase boundaries. Comput. Fluids **75**, 112–126 (2013). https://doi.org/10.1016/j.compfluid.2013.01.024
12. Fechter, S., Munz, C.D.: A discontinuous Galerkin-based sharp-interface method to simulate three-dimensional compressible two-phase flow. Int. J. Numer. Meth. Fluids **78**(7), 413–435 (2015). https://doi.org/10.1002/fld.4022
13. Fedkiw, R.P., Aslam, T., Merriman, B., Osher, S.: A non-oscillatory Eulerian approach to interfaces in multimaterial flows (the ghost fluid method). J. Comput. Phys. **152**(2), 457–492 (1999). https://doi.org/10.1006/jcph.1999.6236
14. Föll, F., Hitz, T., Müller, C., Munz, C.D., Dumbser, M.: On the use of tabulated equations of state for multi-phase simulations in the homogeneous equilibrium limit. Shock Waves (2019). https://doi.org/10.1007/s00193-019-00896-1
15. Fuster, D., Popinet, S.: An all-mach method for the simulation of bubble dynamics problems in the presence of surface tension. J. Comput. Phys. **374**, 752–768 (2018)
16. Grooss, J., Hesthaven, J.: A level set discontinuous Galerkin method for free surface flows. Comput. Methods Appl. Mech. Eng. **195**(25–28), 3406–3429 (2006). https://doi.org/10.1016/j.cma.2005.06.020
17. Hindenlang, F., Gassner, G.J., Altmann, C., Beck, A., Staudenmaier, M., Munz, C.D.: Explicit discontinuous Galerkin methods for unsteady problems. Comput. Fluids **61**, 86–93 (2012). https://doi.org/10.1016/j.compfluid.2012.03.006
18. Huerta, A., Casoni, E., Peraire, J.: A simple shock-capturing technique for high-order discontinuous Galerkin methods. Int. J. Numer. Meth. Fluids **69**(10), 1614–1632 (2011). https://doi.org/10.1002/fld.2654
19. Jiang, G.S., Peng, D.: Weighted ENO schemes for Hamilton-Jacobi equations. SIAM J. Sci. Comput. **21**(6), 2126–2143 (2000). https://doi.org/10.1137/s106482759732455x
20. Jolgam, S., Ballil, A., Nowakowski, A., Nicolleau, F.: On equations of state for simulations of multiphase flows. In: Proceedings of the World Congress on Engineering 2012. 4–6 July 2012. London, UK., vol. 3, pp. 1963–1968 (2012)
21. Kennedy, C.A., Carpenter, M.H.: Additive Runge-Kutta schemes for convection–diffusion–reaction equations. Appl. Numer. Math. **44**(1–2), 139–181 (2003). https://doi.org/10.1016/s0168-9274(02)00138-1
22. Knoll, D.A., Keyes, D.E.: Jacobian-free Newton-Krylov methods: a survey of approaches and applications. J. Comput. Phys. **193**(2), 357–397 (2004)
23. Kopriva, D.A.: Spectral element methods. In: Scientific Computation, pp. 293–354. Springer Netherlands (2009). https://doi.org/10.1007/978-90-481-2261-5_8
24. Krais, N., Beck, A., Bolemann, T., Frank, H., Flad, D., Gassner, G., Hindenlang, F., Hoffmann, M., Kuhn, T., Sonntag, M., Munz, C.D.: FLEXI: A high order discontinuous Galerkin framework for hyperbolic-parabolic conservation laws. Comput. Math. with Appl. (2020). https://doi.org/10.1016/j.camwa.2020.05.004
25. Lervag, K.Y.: Calculation of interface curvature with the level-set method. arXiv preprint arXiv:1407.7340 (2014)
26. Lervag, K.Y., Müller, B., Munkejord, S.T.: Calculation of the interface curvature and normal vector with the level-set method. Comput. Fluids **84**, 218–230 (2013). https://doi.org/10.1016/j.compfluid.2013.06.004

27. Liu, T., Khoo, B., Wang, C.: The ghost fluid method for compressible gas–water simulation. J. Comput. Phys. **204**(1), 193–221 (2005). https://doi.org/10.1016/j.jcp.2004.10.012
28. Liu, T., Khoo, B., Xie, W.: The modified ghost fluid method as applied to extreme fluid-structure interaction in the presence of cavitation. Commun. Comput. Phys. **1**(5), 898–919 (2006)
29. Liu, T.G., Khoo, B.C., Wang, C.W.: The ghost fluid method for compressible gas-water simulation. J. Comput. Phys. **204**(1), 193–221 (2005). https://doi.org/10.1016/j.jcp.2004.10.012
30. Marchandise, E., Remacle, J.F.: A stabilized finite element method using a discontinuous level set approach for solving two phase incompressible flows. J. Comput. Phys. **219**(2), 780–800 (2006). https://doi.org/10.1016/j.jcp.2006.04.015
31. Müller, C., Hitz, T., Jöns, S., Zeifang, J., Chiocchetti, S., Munz, C.D.: Improvement of the level-set ghost-fluid method for the compressible Euler equations. In: Lamanna, G., Tonini, S., Cossali, G.E., Weigand, B. (eds.) Droplet Interaction and Spray Processes. Springer, Heidelberg, Berlin (2020)
32. Nourgaliev, R.R., Dinh, T.N., Theofanous, T.G.: Adaptive characteristics-based matching for compressible multifluid dynamics. J. Comput. Phys. **213**(2), 500–529 (2006)
33. Persson, P.O., Peraire, J.: Sub-cell shock capturing for discontinuous Galerkin methods. In: 44th AIAA Aerospace Sciences Meeting and Exhibit. American Institute of Aeronautics and Astronautics (2006). https://doi.org/10.2514/6.2006-112
34. Saurel, R., Cocchi, J.P., Butler, P.B.: Numerical study of cavitation in the wake of a hypervelocity underwater projectile. J. Propuls. Power **15**(4), 513–522 (1999). https://doi.org/10.2514/2.5473
35. Saurel, R., Petitpas, F., Abgrall, R.: Modelling phase transition in metastableliquids: application to cavitating and flashing flows. J. Fluid Mech. **607** (2008). https://doi.org/10.1017/s0022112008002061
36. Schleper, V.: A HLL-type Riemann solver for two-phase flow with surface forces and phase transitions. Appl. Numer. Math. **108**, 256–270 (2016). https://doi.org/10.1016/j.apnum.2015.12.010
37. Sethian, J.A.: A fast marching level set method for monotonically advancing fronts. Proc. Natl. Acad. Sci. **93**(4), 1591–1595 (1996)
38. Sonntag, M.: Shape derivatives and shock capturing for the Navier-Stokes equations in discontinuous Galerkin methods. Dissertation, University of Stuttgart (2017)
39. Sonntag, M., Munz, C.D.: Efficient parallelization of a shock capturing for discontinuous Galerkin methods using finite volume sub-cells. J. Sci. Comput. **70**(3), 1262–1289 (2016). https://doi.org/10.1007/s10915-016-0287-5
40. Sussman, M., Smereka, P., Osher, S.: A level set approach for computing solutions to incompressible two-phase flow. J. Comput. Phys. **114**(1), 146–159 (1994). https://doi.org/10.1006/jcph.1994.1155
41. Tanguy, S., Berlemont, A.: Application of a level set method for simulation of droplet collisions. Int. J. Multiph. Flow **31**(9), 1015–1035 (2005)
42. Toro, E.F., Spruce, M., Speares, W.: Restoration of the contact surface in the HLL-Riemann solver. Shock Waves **4**(1), 25–34 (1994). https://doi.org/10.1007/bf01414629
43. Tsai, Y.H.R., Cheng, L.T., Osher, S., Zhao, H.K.: Fast sweeping algorithms for a class of Hamilton-Jacobi equations. SIAM J. Numer. Anal. **41**(2), 673–694 (2003)
44. Vahab, M., Miller, G.: A front-tracking shock-capturing method for two gases. Commun. Appl. Math. Comput. Sci. **11**(1), 1–35 (2015)
45. Wang, C.W., Liu, T.G., Khoo, B.C.: A real ghost fluid method for the simulation of multimedium compressible flow. SIAM J. Sci. Comput. **28**(1), 278–302 (2006). https://doi.org/10.1137/030601363
46. Wardlaw Jr., A.B., Luton, J.A.: Fluid-structure interaction mechanisms for close-in explosions. Shock. Vib. **7**(5), 265–275 (2000)
47. Wardlaw Jr., A.B., Mair, H.U.: Spherical solutions of an underwater explosion bubble. Shock. Vib. **5**(2), 89–102 (1998)
48. Williamson, J.H.: Low-storage Runge-Kutta schemes. J. Comput. Phys. **35**, 48–56 (1980)

49. Winter, J.M., Kaiser, J.W., Adami, S., Adams, N.A.: Numerical investigation of 3D drop-breakup mechanisms using a sharp interface level-set method. In: 11th International Symposium Turbulence Shear Flow Phenomena, TSFP 2019 (2001), pp. 1–6 (2019)
50. Zeifang, J., Kaiser, K., Beck, A., Schütz, J., Munz, C.D.: Efficient high-order discontinuous Galerkin computations of low Mach number flows. Commun. Appl. Math. Comput. Sci. **13**(2), 243–270 (2018)
51. Zeifang, J., Schütz, J., Kaiser, K., Beck, A., Lukacova-Medvid'ova, M., Noelle, S.: A novel full-Euler low Mach number IMEX splitting. Commun. Comput. Phys. **27**(1), 292–320 (2019). https://doi.org/10.4208/cicp.OA-2018-0270
52. Zheng, H.W., Shu, C., Chew, Y.T., Qin, N.: A solution adaptive simulation of compressible multi-fluid flows with general equation of state. Int. J. Numer. Methods Fluids **67**(5), 616–637 (2011). https://doi.org/10.1002/fld.2380

Entropy Stable Numerical Fluxes for Compressible Euler Equations Which Are Suitable for All Mach Numbers

Jonas P. Berberich and Christian Klingenberg

Abstract We propose two novel two-state approximate Riemann solvers for the compressible Euler equations which are provably entropy dissipative and suitable for the simulation of low Mach numbers. What is new, is that one of our two methods in addition is provably kinetic energy stable. Both methods are based on the entropy satisfying and kinetic energy consistent methods of [5]. The low Mach number compliance is achieved by rescaling some speed of sound terms in the diffusion matrix in the spirit of [17]. In numerical tests we demonstrate the low Mach number compliance and the entropy stability of the proposed fluxes.

1 Introduction

Compressible Euler equations are used to model the flow of compressible inviscid fluids such as air. To find approximate solutions of the Euler system it is common to use finite volume methods. These are well-suited due to their conservative nature and their capability to resolve discontinuities. The fluxes at the cell interfaces are often determined by approximating the solution of the 1-d interface Riemann problem using numerical (two-state) fluxes.

For a sequence of ever lower Mach numbers, the solutions of the compressible Euler equations with well-prepared initial data converge towards solutions of the incompressible Euler equations [8]. This limit, however, is not correctly represented in a finite volume scheme using conventional numerical fluxes due to excessive diffusion at low Mach numbers. Special low Mach number compliant numerical fluxes have been developed (e.g. [2, 4, 7, 17–19, 22, 27, 28], to correct this behavior.

Stability is required to ensure the convergence of a finite volume method. There are different notions of stability, one of them being entropy stability. Entropy stability is a non-linear stability criterion which additionally ensures that the entropy inequality

J. P. Berberich (✉) · C. Klingenberg
Department of Mathematics, University of Würzburg, Emil-Fischer-Straße 40, 97074 Würzburg, Germany
e-mail: klingen@mathematik.uni-wuerzburg.de

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*,
SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2_8

-an admissibility criterion for physical solutions- is satisfied. Ismail and Roe [14], and later Chandrashekar [5], developed two-state Riemann solvers based on the Roe flux [23], which ensure entropy dissipation to achieve entropy stability. The flux by Chandrashekar [5] is kinetic energy consistent additionally.

Recently, a numerical flux based on [14] has been developed which is entropy dissipative and low Mach compliant [6]. In this article we present two methods for compressible Euler equations closed with an ideal gas law based on the entropy stable and kinetic energy compliant fluxes of Chandrashekar [5] and a low Mach modification of Li and Gu [17]. The methods we propose are entropy dissipative and low Mach compliant. One of our methods is additionally kinetic energy stable. The low Mach number method from Li and Gu [17] has the tendency to develop so-called checkerboard instabilities at low Mach numbers. Numerical experiments indicate that the methods developed in this article suppress the checkerboard modes.

The rest of the article is structured as follows: In Sect. 2 we describe the entropy and kinetic energy dissipative methods introduced in [5]. In Sect. 3 we discuss the behavior of the method at low Mach numbers and we present a correction in Sect. 4. Numerical tests demonstrating the improved results at low Mach numbers and the entropy dissipation of the proposed methods are presented in Sect. 5.

2 Kinetic Energy and Entropy Stable Fluxes

The 2-d Euler equations which model the conservation laws of mass, momentum, and energy of a compressible inviscid fluid are given by

$$\frac{\partial \boldsymbol{q}}{\partial t} + \frac{\partial \boldsymbol{f}}{\partial x} + \frac{\partial \boldsymbol{g}}{\partial y} = 0, \tag{1}$$

where the conserved variables and fluxes are

$$\boldsymbol{q} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ E \end{bmatrix}, \quad \boldsymbol{f} = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ (E+p)u \end{bmatrix}, \quad \boldsymbol{g} = \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ (E+p)v \end{bmatrix}. \tag{2}$$

Moreover, $E = \rho\varepsilon + \frac{1}{2}\rho|\mathbf{v}|^2$ is the total energy per unit volume with $\mathbf{v} = [u, v]^T$ being the velocity. The pressure p is related to the density and internal energy via the ideal gas equation of state

$$p = RT\rho \qquad \text{with} \qquad T = \frac{\gamma - 1}{R}\frac{\varepsilon}{\rho}. \tag{3}$$

The dependent variable T is called temperature, the constants are the gas constant R, and the ratio of specific heats γ.

2.1 Entropy-Entropy Flux and Entropy Variables

A pair $(U, \boldsymbol{\phi})$ with a convex function $U(\boldsymbol{q})$ and a vector valued function $\boldsymbol{\phi}(\boldsymbol{q}) = [\phi_x(\boldsymbol{q}), \phi_y(\boldsymbol{q})]^T$ is called *entropy-entropy flux pair*, if it satisfies the relations

$$U'(\boldsymbol{q})\boldsymbol{f}'(\boldsymbol{q}) = \phi_x'(\boldsymbol{q}), \qquad U'(\boldsymbol{q})\boldsymbol{g}'(\boldsymbol{q}) = \phi_y'(\boldsymbol{q}). \tag{4}$$

Using this pair, we can add the additional conservation law

$$\frac{\partial U}{\partial t} + \frac{\partial \phi_x}{\partial x} + \frac{\partial \phi_y}{\partial y} = 0 \tag{5}$$

to Eq. (1). As usual in the context of hyperbolic conservation laws, we also want to admit discontinuous solutions and interpret all the derivatives in Eq. (1) in the weak sense. At discontinuities, the entropy is not necessarily conserved. Instead, the inequality

$$\frac{\partial U}{\partial t} + \frac{\partial \phi_x}{\partial x} + \frac{\partial \phi_y}{\partial y} \leq 0 \tag{6}$$

is demanded as a criterion to choose admissible (physical) solutions. We define *entropy variables* by

$$\mathbf{r}(\boldsymbol{q}) := U'(\boldsymbol{q}) \tag{7}$$

and the dual to the entropy flux by $\boldsymbol{\psi}(\mathbf{r}) = [\psi_x(\mathbf{r}), \psi_y(\mathbf{r})]^T$ with

$$\psi_x(\mathbf{r}) := \mathbf{r} \cdot \boldsymbol{f}(\boldsymbol{q}(\mathbf{r})) - \phi_x(\boldsymbol{q}(\mathbf{r})), \quad \psi_y(\mathbf{r}) := \mathbf{r} \cdot \boldsymbol{g}(\boldsymbol{q}(\mathbf{r})) - \phi_y(\boldsymbol{q}(\mathbf{r})), \tag{8}$$

where $\boldsymbol{q}(\mathbf{r})$ is the inverse of $\mathbf{r}(\boldsymbol{q})$ defined above. The inverse exists because of the convexity of $U(\boldsymbol{q})$.

For the Euler Eq. (1), the most common choice of an entropy-entropy flux pair is

$$U := -\frac{\rho s}{\gamma - 1}, \qquad \boldsymbol{\phi} := -\frac{\rho \mathbf{v} s}{\gamma - 1}, \tag{9}$$

where $s := \ln\left(p\rho^{-\gamma}\right) = -(\gamma - 1)\ln(\rho) - \ln(\beta) - \ln(2)$ up to a constant with $\beta := 1/(2RT)$. This choice is not unique [12], but it is the one consistent with the entropy condition from thermodynamics in the presence of heat transfer [13]. The entropy variables are subsequently given by

$$\mathbf{r} := \left[\tfrac{[\gamma - s]}{\gamma - 1} - \beta|\mathbf{v}|^2,\ 2\beta u,\ 2\beta v,\ -2\beta\right]^T. \tag{10}$$

and the entropy flux dual

$$\boldsymbol{\psi} = \rho \mathbf{v}. \tag{11}$$

2.2 A Basic Finite Volume Method

In this section, for brevity reasons, we only describe a simple quadrature-free finite volume method, which is a second order accurate finite volume method on a static Cartesian grid. In practice, more elaborate methods are used (see e.g. [26]).

We divide the domain $\Omega = [a, b] \times [c, d]$ with $a < b$, $c < d$ into cells

$$\Omega_{ij} := \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right] \times \left[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}\right] \tag{12}$$

for $i = 0, \ldots, N-1$, $j = 0, \ldots, M-1$. The cell-interface centers are

$$\mathbf{x}_{i-\frac{1}{2},j} := \left[a + i\Delta x, c + \left(j + \frac{1}{2}\right)\Delta y\right]^T, \mathbf{x}_{i,j-\frac{1}{2}} := \left[a + \left(i + \frac{1}{2}\right)\Delta x, c + j\Delta y\right]^T \tag{13}$$

with $\Delta x := \frac{b-a}{N}$, $\Delta y := \frac{d-c}{M}$. We integrate $\boldsymbol{q}$ in each cell to obtain cell-averaged values

$$\hat{\boldsymbol{q}}_{ij}(t) := \frac{1}{\Delta x \Delta y} \int_{\Omega_{ij}} \boldsymbol{q}(\mathbf{x}, t)\, d\mathbf{x}. \tag{14}$$

We find an evolution equation for the cell-average values by cell-wise integrating Eq. (1):

$$\partial_t \hat{\boldsymbol{q}}_{ij}(t) = -\frac{1}{\Delta x \Delta y} \int_{\Omega_{ij}} \partial_x \boldsymbol{f}(\boldsymbol{q}(\mathbf{x}, t)) + \partial_y \boldsymbol{g}(\boldsymbol{q}(\mathbf{x}, t))\, d\mathbf{x}. \tag{15}$$

To construct our simple finite volume method we use Fubini's theorem, the fundamental theorem of calculus, and an approximation of the interface integral by the interface centered point value. The interface centered fluxes are approximated using a numerical two-state flux. This yields

$$\begin{aligned} \partial_t \hat{\boldsymbol{q}}_{ij}(t) \approx &- \frac{1}{\Delta x}\left[\mathbf{F}\left(\hat{\boldsymbol{q}}^-_{i+\frac{1}{2},j}(t), \hat{\boldsymbol{q}}^+_{i+\frac{1}{2},j}(t)\right) - \mathbf{F}\left(\hat{\boldsymbol{q}}^-_{i-\frac{1}{2},j}(t), \hat{\boldsymbol{q}}^+_{i-\frac{1}{2},j}(t)\right)\right] \\ &- \frac{1}{\Delta y}\left[\mathbf{G}\left(\hat{\boldsymbol{q}}^-_{i,j+\frac{1}{2}}(t), \hat{\boldsymbol{q}}^+_{i,j+\frac{1}{2}}(t)\right) - \mathbf{G}\left(\hat{\boldsymbol{q}}^-_{i,j-\frac{1}{2}}(t), \hat{\boldsymbol{q}}^+_{i,j-\frac{1}{2}}(t)\right)\right], \end{aligned} \tag{16}$$

where the $\hat{\boldsymbol{q}}^{\pm}$ values are obtained using a non-oscillatory reconstruction on the cell-average values. This set of ODEs is then integrated numerically to evolve the approximate solution $\hat{\boldsymbol{q}}$ in time. In the rest of the article we drop the hat at $\hat{\boldsymbol{q}}_{ij}$ and just write $\boldsymbol{q}_{ij}$. Also, in the rest of the article we will only consider numerical fluxes in x-direction, since the fluxes $\boldsymbol{f}$ and $\boldsymbol{g}$ for Euler equations can be converted into each

other by only correctly rotating velocity vectors. For symmetry reasons we assume $\mathbf{F}$ and $\mathbf{G}$ to have the same relation.

2.3 Entropy Conservative Numerical Fluxes

Tadmor [24, 25] introduced the concept of *entropy conservative numerical fluxes* $\mathbf{F}^{\mathrm{ec}}$, which have to satisfy the relation

$$(\mathbf{r}(\boldsymbol{q}^+) - \mathbf{r}(\boldsymbol{q}^-)) \cdot \mathbf{F}^{\mathrm{ec}}(\boldsymbol{q}^-, \boldsymbol{q}^+) = \psi(\mathbf{r}(\boldsymbol{q}^+)) - \psi(\mathbf{r}(\boldsymbol{q}^-)) = (\rho\mathbf{v})^+ - (\rho\mathbf{v})^-. \quad (17)$$

The last identity is only valid for the Euler Eq. (1). Different entropy conservative numerical fluxes have been proposed by Tadmor [24], Ismail and Roe [14], and Chandrashekar [5]. Our method is based on the numerical flux by Chandrashekar [5], which can be written as

$$\mathbf{F}^*(\boldsymbol{q}^-, \boldsymbol{q}^+) := \begin{bmatrix} F^{*,\rho} \\ F^{*,\rho u} \\ F^{*,\rho v} \\ F^{*,E} \end{bmatrix} := \begin{bmatrix} \hat{\rho}\bar{u} \\ \bar{u}F^{*,\rho} + \tilde{p} \\ \bar{v}F^{*,\rho} \\ \left(\frac{1}{2(\gamma-1)\tilde{\beta}} - \frac{1}{2}\overline{|\mathbf{v}|^2}\right) F^{*,\rho} + \bar{\mathbf{v}} \cdot [F^{*,\rho u}, F^{*,\rho v}]^T \end{bmatrix}. \quad (18)$$

The averages are the arithmetic average $\bar{a} := \frac{1}{2}(a^- + a^+)$ and the logarithmic average $\hat{a} := \frac{a^+ - a^-}{\ln a^+ - \ln a^-}$. A non-singular implementation of $\hat{a}$ is presented in [14]. The pressure average is $\tilde{p} := \bar{\rho}/(2\bar{\beta})$ where $\bar{\beta}$ is computed from $\beta^\pm = \rho^\pm/(2p^\pm)$. This pressure average corresponds to the harmonic average in the temperature [5]. The notations for the averages are used throughout this article.

2.4 Kinetic Energy Preserving Fluxes

From the density and momentum equations in the Euler Eq. (1) the balance law

$$\frac{\partial K}{\partial t} + \frac{\partial (Ku)}{\partial x} + \frac{\partial (Kv)}{\partial y} = -u\frac{\partial p}{\partial x} - v\frac{\partial p}{\partial y} \quad (19)$$

for the kinetic energy $K := \frac{1}{2}\rho|\mathbf{v}|^2$ can be derived. Integration over the whole domain Ω while ignoring the boundaries yields

$$\frac{\partial}{\partial t}\int_\Omega K \, d\mathbf{x} = \int_\Omega p\frac{\partial u}{\partial x} + p\frac{\partial v}{\partial y} \, d\mathbf{x}. \quad (20)$$

Jameson [15] shows that any numerical flux $\mathbf{F}^J$ which can be formulated in the form

$$\mathbf{F}^{\mathrm{J}}\left(\boldsymbol{q}^{-},\boldsymbol{q}^{+}\right)=\begin{bmatrix} F^{\mathrm{J},\rho} \\ \bar{u}F^{\mathrm{J},\rho}+\langle p\rangle \\ \bar{v}F^{\mathrm{J},\rho} \\ F^{\mathrm{J},E} \end{bmatrix}, \tag{21}$$

satisfies the discrete analogon

$$\begin{aligned}\frac{\partial}{\partial t}\sum_{i,j}K_{ij}\Delta x\Delta y&=\sum_{i,j}\left(-\frac{1}{2}|\mathbf{v}_{ij}|^2\frac{\partial\rho_{ij}}{\partial t}+\mathbf{v}_{ij}\frac{\partial(\rho\mathbf{v})_{ij}}{\partial t}\right)\Delta x\Delta y\\&=\sum_{i,j}\left(\langle p\rangle_{i+\frac{1}{2},j}\frac{\Delta u_{i+\frac{1}{2},j}}{\Delta x}+\langle p\rangle_{i,j+\frac{1}{2}}\frac{\Delta v_{i,j+\frac{1}{2}}}{\Delta y}\right)\Delta x\Delta y\end{aligned} \tag{22}$$

of Eq. (20). Equation (22) can easily be computed using the flux from Eq. (21) and the corresponding flux in y-direction $\mathbf{G}^{J}$. The fluxes $F^{\mathrm{J},\rho}$ and $F^{\mathrm{J},E}$ are consistent approximations of the density and energy flux and $\langle p\rangle$ approximates the interface pressure. Clearly, the entropy conservative numerical flux $\mathbf{F}^{*}$ from Eq. (18) is in this kinetic energy preserving form.

In the low Mach number limit, the right-hand side of Eq. (20) vanishes with the divergence of velocity (e.g. [11]). Equation (20) then describes the conservation of kinetic energy. This makes kinetic energy consistency especially relevant for low Mach number fluxes. Most numerical flux functions violate this condition. For example, in [18] it is numerically shown that the kinetic energy rises for a simulation of the incompressible Gresho [10] vortex using a central flux and implicit time stepping.

2.5 *Entropy Diffusion*

For the scheme to be stable in the presence of discontinuities it needs to dissipate entropy. Following [5] this is achieved by modifying the Roe scheme diffusion [23] such that the Roe matrix is applied to the jump in entropy variables $\mathbf{r}$ instead of conserved variables $\boldsymbol{q}$. The standard Roe scheme uses the diffusion

$$\mathbf{F}^{\mathrm{Roe,diff}}\left(\boldsymbol{q}^{-},\boldsymbol{q}^{+}\right):=-\frac{1}{2}D^{\mathrm{Roe}}\Delta\boldsymbol{q}:=-\frac{1}{2}R|\Lambda|^{\mathrm{Roe}}R^{-1}\Delta\boldsymbol{q}, \tag{23}$$

where

$$R:=\begin{bmatrix} 1 & 1 & 0 & 1 \\ u-c & u & 0 & u+c \\ v & v & -1 & v \\ H-cu & \frac{1}{2}|\mathbf{v}|^2 & -v & H+cu \end{bmatrix}, \tag{24}$$

with the enthalpy $H=\frac{c^2}{\gamma-1}+\frac{|\mathbf{v}|^2}{2}$, is the matrix of right eigenvectors of $\frac{\partial \boldsymbol{f}(\boldsymbol{q})}{\partial \boldsymbol{q}}$ and

$$|\Lambda|^{\text{Roe}} := \text{diag}\left(\left[|\lambda_1|, |\lambda_2|, |\lambda_3|, |\lambda_4|\right]\right) = \text{diag}\left(\left[|u-c|, |u|, |u|, |u+c|\right]\right) \quad (25)$$

is the diagonal matrix with the absolute values of the corresponding eigenvalues. The whole matrix D^{Roe} is evaluated at the Roe average state [23] to ensure accurate shock capturing. In order to apply the diffusion matrix D^{Roe} to the jump in entropy variables, we have to transform these to conserved variables. So the diffusion part of our numerical flux is

$$\mathbf{F}^{\text{ES,diff}}\left(\boldsymbol{q}^-, \boldsymbol{q}^+\right) := -\frac{1}{2} R|\Lambda|^{\text{Roe}} R^{-1} \frac{\partial \boldsymbol{q}}{\partial \mathbf{r}} \Delta \mathbf{r}. \quad (26)$$

The entropy diffusion $\mathbf{F}^{\text{ES,diff}}$ can be formulated in a simpler form which can lead to a more efficient implementation: It is shown by Barth [3] that there is a scaling $\tilde{R} = RS^{-\frac{1}{2}}$ of the Eigenvectors R with $\frac{\partial \boldsymbol{q}}{\partial \mathbf{r}} = \tilde{R}\tilde{R}^T$ which leads to the form

$$\mathbf{F}^{\text{ES,diff}}\left(\boldsymbol{q}^-, \boldsymbol{q}^+\right) = -\frac{1}{2}\tilde{R}|\Lambda|^{\text{Roe}}\tilde{R}^{-1}\tilde{R}\tilde{R}^T\Delta\mathbf{r} = -\frac{1}{2}\underbrace{R|\Lambda|^{\text{Roe}}SR^T}_{=:Q}\Delta\mathbf{r} = -\frac{1}{2}Q\Delta\mathbf{r} \quad (27)$$

with the scaling matrix

$$S := \text{diag}\left(\left[\tfrac{\rho}{2\gamma}, \tfrac{(\gamma-1)\rho}{\gamma}, p, \tfrac{\rho}{2\gamma}\right]\right). \quad (28)$$

Since Q is positive definite by construction, $\mathbf{F}^{\text{ES,diff}}$ is dissipative in the entropy variables. The numerical flux

$$\mathbf{F}^{\text{ES}}\left(\boldsymbol{q}^-, \boldsymbol{q}^+\right) := \mathbf{F}^*\left(\boldsymbol{q}^-, \boldsymbol{q}^+\right) + \mathbf{F}^{\text{ES,diff}}\left(\boldsymbol{q}^-, \boldsymbol{q}^+\right) \quad (29)$$

is hence entropy satisfying in the sense, that a spatially discrete analogon of the entropy inequality is satisfied. Proofs are analogously to [5, 20]. The numerical flux in y-direction is constructed in the same way.

2.6 Kinetic Energy Diffusion

Alongside entropy stability we also aim for kinetic energy stability. We have seen that $\mathbf{F}^*$ ist consistent with the evolution of kinetic energy derived from the Euler Eq. (1). In order to guarantee kinetic energy stability, the diffusive term in the numerical flux has to do dissipate kinetic energy. Chandrashekar [5] shows that this requires the condition $|\lambda_1| = |\lambda_4|$ to hold in Eq. (25). The most obvious way to achieve this is to choose

$$|\Lambda|^{\text{KES}} := \text{diag}\left(\left[|\lambda_1|, |\lambda_2|, |\lambda_3|, |\lambda_4|\right]\right) = \text{diag}\left(\left[|u|+c, |u|, |u|, |u|+c\right]\right). \quad (30)$$

in the definition of the entropy-diffusion matrix Q. The numerical flux defined by this modification on the ES scheme will be called ES-KES or $\mathbf{F}^{\text{ES-KES}}$ throughout this article. Note that this method is more diffusive than the ES scheme. However, it adds one more relevant stability property.

2.7 Intermediate State

To ensure correct upwinding, the diffusion matrix in the standard Roe scheme is evaluated at the so-called Roe average state. Anyway, in our numerical flux $\mathbf{F}^{\text{ES}}$ we use $\mathbf{F}^*$ instead of the standard central flux, so we can not expect the shock-capturing property to still hold for our method. In the following we discuss at which intermediate state the diffusion matrix Q should be evaluated.

The entropy stability property does not depend on the intermediate state, since Q is positive definite for any state by construction. The kinetic energy stability also does not depend on the particular choice of the intermediate state, only on the relation of the entries in the diagonal matrix $|\Lambda|$. For the flux to have the contact property, Chandrashekar [5] shows that we have to choose

$$c_{\text{int}} = \sqrt{\frac{\gamma}{2\hat{\beta}}} \quad \text{and} \quad H_{\text{int}} = \frac{c_{\text{int}}^2}{\gamma - 1} + \frac{1}{2}|\mathbf{v}_{\text{int}}|^2. \tag{31}$$

All other averages can be chosen freely, so we can use arithmetic or logarithmic averages for example.

However, for implementation reasons it can be useful to hand a intermediate state vector in primitive variables to the routine which computes the diffusion matrix. We can realize this using the average state

$$\mathbf{v}_{\text{int}} := \bar{\mathbf{v}} \qquad p_{\text{int}} := \bar{p} \qquad \rho_{\text{int}} := 2p_{\text{int}}\hat{\beta}. \tag{32}$$

In the computation of the diffusion matrix we compute all the other variables from the primitive intermediate state in the straight forward way, e.g. $c_{\text{int}} = \sqrt{\gamma p_{\text{int}}/\rho_{\text{int}}}$ and the enthalpy as described in Eq. (31).

3 Low Mach Number Asymptotics

The Euler Eq. (1) can be cast in the non-dimensional form

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho\\ \rho u\\ \rho v\\ E\end{bmatrix}+\frac{\partial}{\partial x}\begin{bmatrix}\rho u\\ \rho u^2+\frac{1}{\mathcal{M}^2}p\\ \rho uv\\ (E+p)u\end{bmatrix}+\frac{\partial}{\partial y}\begin{bmatrix}\rho v\\ \rho uv\\ \rho v^2+\frac{1}{\mathcal{M}^2}p\\ (E+p)v\end{bmatrix}=0 \tag{33}$$

using only the assumption that the reference velocity is computed as the quotient of the reference length and time and one parameter, which we will call reference Mach number $\mathcal{M}$.

The low Mach number limit of the Euler equations Eq. (33) is well-known and studied (e.g. [1, 8, 11]). For well-prepared initial data a series of solutions of Eq. (33) with different reference Mach numbers converge to solutions of the incompressible Euler equations for $\mathcal{M}\to 0$. Conventional finite volume methods tend to fail to correctly represent this limit for their numerical solutions. One reason for this is excessive diffusion at low Mach numbers.

Consider the ES-flux introduced in Sect. 2.5. Note that, again, we only consider the flux in x-direction for simplicity. For small jumps we can approximate

$$\begin{aligned}\mathbf{F}^{\mathrm{ES}}\left(\boldsymbol{q}^-,\boldsymbol{q}^+\right)&=\mathbf{F}^*\left(\boldsymbol{q}^-,\boldsymbol{q}^+\right)-\frac{1}{2}D^{\mathrm{Roe}}(\boldsymbol{q}_{\mathrm{int}})\left.\frac{\partial \boldsymbol{q}(\mathbf{r})}{\partial \mathbf{r}}\right|_{\mathbf{r}=\mathbf{r}(\boldsymbol{q}_{\mathrm{int}})}\left(\mathbf{r}\left(\boldsymbol{q}^+\right)-\mathbf{r}\left(\boldsymbol{q}^-\right)\right)\\ &\approx J\left(\boldsymbol{q}_{\mathrm{int}}\right)\boldsymbol{q}_{\mathrm{int}}-\frac{1}{2}D^{\mathrm{Roe}}\left(\boldsymbol{q}_{\mathrm{int}}\right)\Delta\boldsymbol{q}\\ &=\left.\frac{\partial \boldsymbol{q}}{\partial \mathbf{u}}\right|_{\boldsymbol{q}_{\mathrm{int}}}J_{\mathrm{prim}}\left(\boldsymbol{q}_{\mathrm{int}}\right)\left.\frac{\partial \mathbf{u}}{\partial \boldsymbol{q}}\right|_{\boldsymbol{q}_{\mathrm{int}}}\boldsymbol{q}_{\mathrm{int}}-\frac{1}{2}\left.\frac{\partial \boldsymbol{q}}{\partial \mathbf{u}}\right|_{\boldsymbol{q}_{\mathrm{int}}}D^{\mathrm{Roe}}_{\mathrm{prim}}\left(\boldsymbol{q}_{\mathrm{int}}\right)\left.\frac{\partial \mathbf{u}}{\partial \boldsymbol{q}}\right|_{\boldsymbol{q}_{\mathrm{int}}}\Delta\boldsymbol{q}\end{aligned} \tag{34}$$

because of $\boldsymbol{q}^-\approx\boldsymbol{q}_{\mathrm{int}}\approx\boldsymbol{q}^+$ and consequently $\mathbf{F}^*\left(\boldsymbol{q}^-,\boldsymbol{q}^+\right)\approx\boldsymbol{f}(\boldsymbol{q}_{\mathrm{int}})$. The intermediate state $\boldsymbol{q}_{\mathrm{int}}$ is the one defined in Sect. 2.7. The flux Jacobian in primitive variables is $J_{\mathrm{prim}}=\frac{\partial \mathbf{u}}{\partial \boldsymbol{q}}J\frac{\partial \boldsymbol{q}}{\partial \mathbf{u}}$ with the flux jacobian in conserved variables $J=\frac{\partial f}{\partial \boldsymbol{q}}$ and the Roe diffusion matrix in primitive variables is $D^{\mathrm{Roe}}_{\mathrm{prim}}=\frac{\partial \mathbf{u}}{\partial \boldsymbol{q}}D^{\mathrm{Roe}}\frac{\partial \boldsymbol{q}}{\partial \mathbf{u}}$. The primitive variables are

$$\mathbf{u}:=[\rho,u,v,p]^T. \tag{35}$$

Equation (34) justifies the comparison of the two matrices J_{prim} and $D^{\mathrm{Roe}}_{\mathrm{prim}}$ with regard of their formal scaling with the reference Mach number to gain insight into the asymptotic behavior of the method for small Mach numbers:

$$J_{\text{prim}} = \begin{bmatrix} \mathcal{O}(1) & \mathcal{O}(1) & 0 & 0 \\ 0 & \mathcal{O}(1) & 0 & \mathcal{O}\left(\frac{1}{\mathcal{M}^2}\right) \\ 0 & 0 & \mathcal{O}(1) & 0 \\ 0 & \mathcal{O}(1) & 0 & \mathcal{O}(1) \end{bmatrix}, \tag{36}$$

$$D_{\text{prim}}^{\text{Roe}} = \begin{bmatrix} \mathcal{O}(1) & 0 & 0 & \mathcal{O}\left(\frac{1}{\mathcal{M}}\right) \\ 0 & \mathcal{O}\left(\frac{1}{\mathcal{M}}\right) & 0 & \mathcal{O}\left(\frac{1}{\mathcal{M}}\right) \\ 0 & 0 & \mathcal{O}(1) & \mathcal{O}(1) \\ 0 & 0 & 0 & \mathcal{O}\left(\frac{1}{\mathcal{M}}\right) \end{bmatrix} + \mathcal{O}(\mathcal{M}). \tag{37}$$

Note that the diffusion matrix scaling for the ES-KES flux is the same, since for low Mach numbers $|u + c| \approx |u| + c \approx |u - c|$. From Eqs. (36) and (37) we see that there are some terms in which the diffusion matrix dominates the flux Jacobian for small Mach numbers. This explains the excessive diffusion of the method at low Mach numbers. Different methods have been proposed to correct those terms in the Roe diffusion matrix (e.g. [2, 18, 19, 22, 27]). For the entropy stability of our method, however, the modification should keep the positive definiteness of Q. The following modification provides this.

4 Low Mach Modifications of the ES and ES-KES Fluxes

Following [17] we modify the diagonal matrices Eqs. (25) and (30) to make the schemes low Mach number compliant: We use

$$|\Lambda|_{\text{LM}}^{\text{Roe}} := \text{diag}\left(\left[|u - \tilde{c}|, |u|, |u|, |u + \tilde{c}|\right]\right), \tag{38}$$

$$|\Lambda|_{\text{LM}}^{\text{KES}} := \text{diag}\left(\left[|u| + \tilde{c}, |u|, |u|, |u| + \tilde{c}\right]\right) \tag{39}$$

with

$$\tilde{c} := c \cdot \max(\min(M, 1), M_{\text{cut}}) \quad \text{with} \quad M := \frac{|\mathbf{v}|}{c} \quad \text{and} \quad M_{\text{cut}} \in [0, 1]. \tag{40}$$

Other possible definitions of the rescaled speed of sound $\tilde{c}$ can be found e.g. in [6, 17]. The global cut-off Mach number M_{cut} can be used to increase the diffusion and thus the stability at Mach numbers which are lower than the one expected in a particular simulation. Using those diagonal matrices instead of $|\Lambda|^{\text{Roe}}$ and $|\Lambda|^{\text{KES}}$ yields the numerical fluxes $\mathbf{F}^{\text{ES-LM}}$ and $\mathbf{F}^{\text{ES-KES-LM}}$. It is obvious that this modification is compatible with the proof of entropy stability in [5]. Also, $\mathbf{F}^{\text{ES-KES-LM}}$ still satisfies the relation $|\lambda_1| = |\lambda_4|$ required for kinetic energy stability. This modification changes Eq. (37) into

$$\left(D_{\mathrm{LM}}^{\mathrm{Roe}}\right)_{\mathrm{prim}} = \begin{bmatrix} O(1) & 0 & 0 & O(1) \\ 0 & O(1) & 0 & O\left(\frac{1}{M}\right) \\ 0 & 0 & O(1) & 0 \\ 0 & 0 & 0 & O(1) \end{bmatrix} + O(M). \tag{41}$$

and the terms of dominating diffusion at low Mach numbers vanish. As before, this scaling is also valid for the diffusion matrix of the ES-KES-LM flux.

5 Numerical Tests

In our tests we include the standard Roe flux [23] (Roe), the Roe flux with the low Mach fix [17] described in Eq. (38) (Roe-LM), the entropy stable fluxes (ES, ES-KES) from [5] and the entropy stable fluxes with low Mach fix (ES-LM, ES-KES-LM) proposed in this article. In the low Mach methods we use $M_{\mathrm{cut}} = 0$. For time-stepping we use the third order accurate four stage Runge–Kutta Method as described in [16]. In practice, an implicit method should be used to evolve low Mach number flows in time due to the stiffness in time [18]. However, this is not in the scope of this short article.

For the equation of state we choose $\gamma = 1.4$, which is a suitable value to describe air. We set the value of the gas constant to $R = 1$, such that density and pressure have the same order of magnitude.

5.1 Test of the Entropy Stability

One test for the entropy compliance of a method is a standing sound wave, for which e.g. the Roe scheme is well-known to produce a non-physical jump. We solve the 1-d Riemann-problem given by

$$(\rho, u, p)(x < 0.5) := (1, 0.75, 1) \tag{42}$$

$$(\rho, u, p)(x \geq 0.5) := (0.125, 0, 0.1) \tag{43}$$

on 100 equidistant grid cells in the domain $\Omega = [0, 1]$. We use the standard Roe-flux, the standard Roe-flux with the low Mach modification Eq. (38) (Roe-LM), and the entropy stable low Mach methods ES-LM and ES-KES-LM introduced in this article (Sect. 4) with constant (which means no) reconstruction. The result at $t = 0.2$ is shown in Fig. 1. As a reference solution we use a simulation with the local Lax–Friedrichs numerical flux on 100 000 grid cells. We see the non-entropic jump which is produced by the Roe scheme. The low Mach modification in Roe-LM even increases the non-entropic jump. As expected, for the entropy stable all Mach methods ES-LM and ES-KES-LM there is no significant non-entropic jump. It is

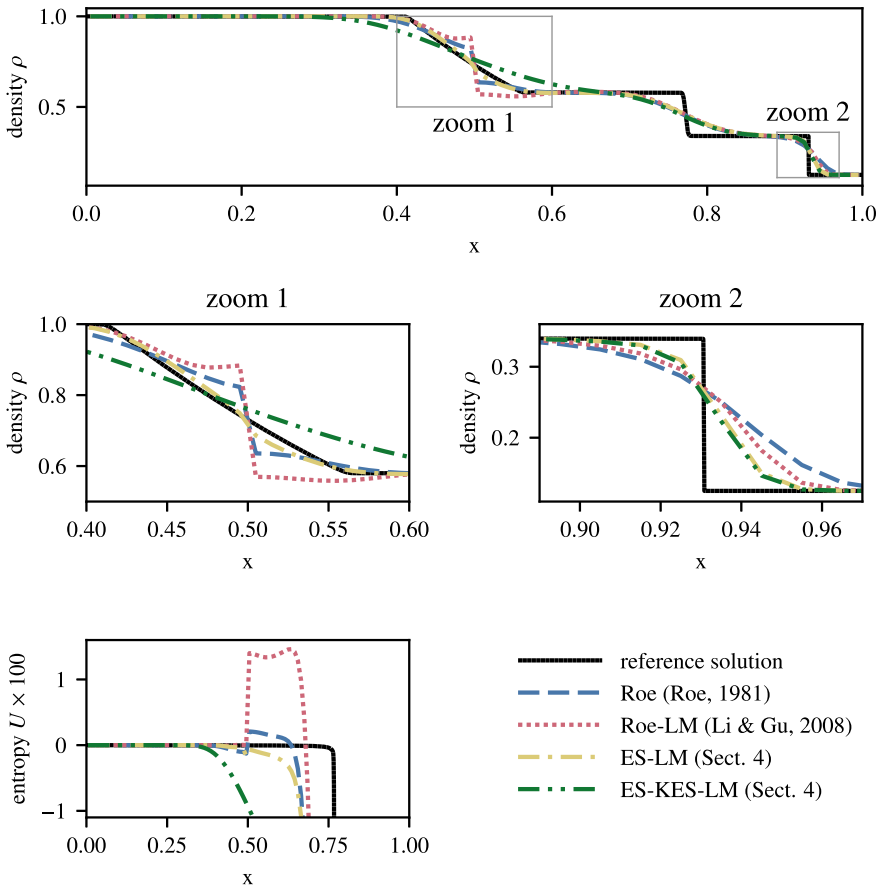


Fig. 1 Standing sound wave test at $t = 0.2$ with different numerical fluxes at a 100 cells grid. The test-setup is described in Sect. 5.1. Top: density on the whole domain. Middle: Magnified views of particular regions of the top panel. Bottom: Entropy at the whole domain

notable that the methods proposed in this article also have an improved accuracy on the expansion shock (zoom 2). In the bottom panel of Fig. 1 we see the entropy of the solution at time $t = 0.2$. While the entropy at initial time is non-positive on the whole domain (This is easy to compute from the initial conditions), the Roe and Roe-LM method lead to positive values of entropy at time $t = 0.2$. The ES-LM and ES-KES-LM method lead to non-positive entropy values on the whole domain.

5.2 Test of the Contact Property

We test the contact property of the method using the Riemann problem

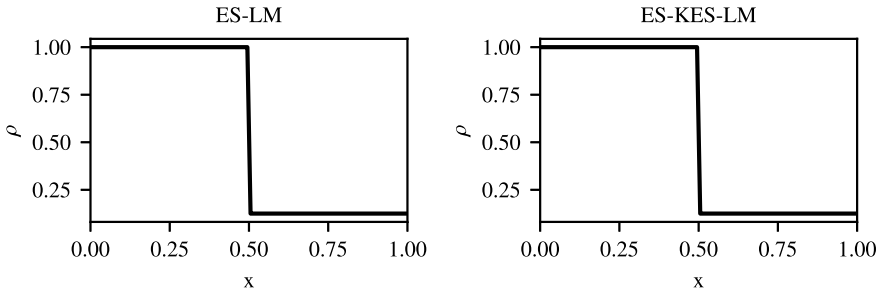


Fig. 2 Density of a contact discontinuity at $t = 0.2$ computed using the entropy stable low Mach methods ES-LM and ES-KES-LM. The setup is described in Sect. 5.2

$$(\rho, u, p)(x < 0.5) := (1, 0.75, 1) \tag{44}$$

$$(\rho, u, p)(x \geq 0.5) := (0.125, 0, 0.1) \tag{45}$$

on 100 equidistant grid cells in the domain [0, 1]. We use the ES-LM and ES-KES-LM fluxes with constant reconstruction and the intermediate state described in Sect. 2.7. The result at $t = 0.2$ is shown in Fig. 2. Both methods accurately resolve the contact discontinuity.

5.3 *Low Mach Gresho Vortex*

The incompressible Gresho vortex [10] can be extended to a family of stationary solutions of the Euler equations with a parameter that adjusts the maximal local Mach number in the setup [2]. The setup is

$$(\rho, \mathbf{v}, p) = \begin{cases} (1,\ 5r\mathbf{e}_\phi,\ p_c + \frac{25}{2}r^2) & \text{if } r < 0.2, \\ \left(1,\ (2-5r)\mathbf{e}_\phi,\ p_c + 4\ln(5r) + 4 - 20r + \frac{25}{2}r^2\right) & \text{if } 0.2 \leq r < 0.4, \\ (1,\ 0,\ p_c + 4\ln(2) - 2) & \text{else,} \end{cases}$$

where $p_c = \frac{1}{\gamma \tilde{M}^2}\frac{1}{2}$ and $\mathbf{e}_\phi$ is the unit vector in angular direction. We apply the standard Roe flux, the Roe flux with low Mach fix [17], the entropy stable fluxes from [5] and the entropy stable fluxes with low Mach fix proposed in this article. We use limited linear reconstruction on a 32×32 cells grid with periodic boundary conditions to evolve the vortex for 0.1 revolutions. For the low Mach number fluxes we use $M_{\text{cut}} = 0$. The Mach number at final time is shown in Fig. 3 for the different numerical fluxes (columns) and maximal initial Mach number parameters $\tilde{M} = 1, 0.1, 0.01$ (rows). The Roe, ES, and ES-KES numerical fluxes lead to completely diffused vortexes for lower Mach number, while the proposed fluxes (ES-LM, ES-KES-LM) are as

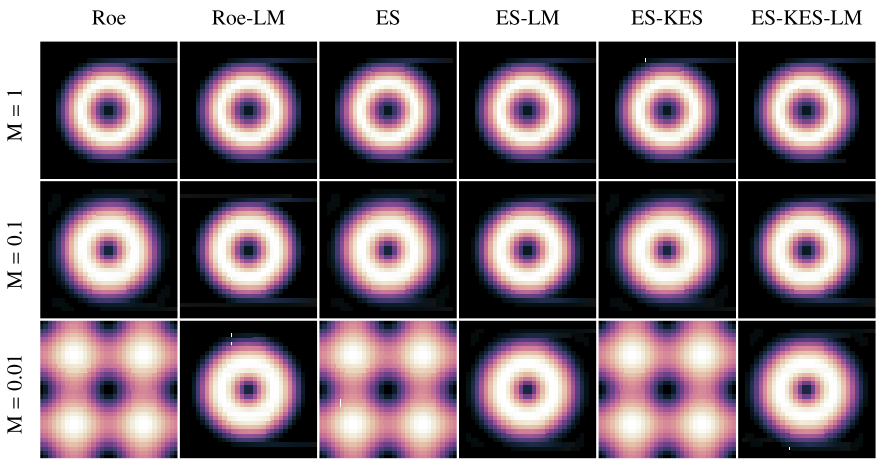


Fig. 3 Local Mach number of the Gresho vortex test from Sect. 5.3 after 0.1 rotations. The initial maximal Mach number decreases from top to bottom. In the different columns different numerical fluxes are applied

capable of accurately resolving the Gresho vortex at lower Mach numbers as the Roe method with low Mach fix (Roe-LM, [17]).

6 Conclusions and Outlook

We presented novel numerical flux functions which combine the entropy and kinetic energy stability properties of the fluxes proposed in [5] with the low Mach number compliance of the method from [17]. The entropy stability and the low Mach number compliance have been shown in numerical tests. The contact property holds due to the correct choice of the intermediate state. It is worth noting that the proposed methods also show an improved performance at fast shocks. For practical applications of the proposed methods we suggest the combination with implicit time-stepping to overcome the stiffness in time. To extend the method to higher order while maintaining the entropy stability, special care has to be given to the reconstruction procedure, as discussed in [9, 21]. To avoid the carbuncle phenomenon at strong shocks, [5, 20] suggest a hybrid diffusion with a Rusanov-type diffusion term.

Acknowledgments The authors thankfully acknowledge the helpful discussions with Praveen Chandrashekar. Jonas Berberich thanks the organizers of the NumHyp 2019 for a science-wise interesting and socially pleasurable conference. The research of Jonas Berberich is supported by the Klaus Tschira Foundation.

References

1. Barsukow, W., Edelmann, P.V., Klingenberg, C., Röpke, F.K.: A low-Mach Roe-type solver for the Euler equations allowing for gravity source terms. ESAIM: Proc. Surv. **58**, 27–39 (2017)
2. Barsukow, W., Edelmann, P.V.F., Klingenberg, C., Miczek, F., Röpke, F.K.: A numerical scheme for the compressible low-Mach number regime of ideal fluid dynamics. J. Sci. Comput. 1–24 (2017). https://doi.org/10.1007/s10915-017-0372-4
3. Barth, T.J.: Numerical methods for gasdynamic systems on unstructured meshes. In: An Introduction to Recent Developments in Theory and Numerics for Conservation Laws, pp. 195–285. Springer (1999)
4. Chalons, C., Girardin, M., Kokh, S.: An all-regime Lagrange-projection like scheme for the gas dynamics equations on unstructured meshes. Commun. Comput. Phys. **20**(1), 188–233 (2016)
5. Chandrashekar, P.: Kinetic energy preserving and entropy stable finite volume schemes for compressible Euler and Navier-Stokes equations. Commun. Comput. Phys. **14**(5), 1252–1286 (2013)
6. Chen, S.S., Yan, C., Lou, S., Lin, B.X.: An improved entropy-consistent Euler flux in low Mach number. J. Comput. Sci. **27**, 271–283 (2018)
7. Dellacherie, S.: Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number. J. Comput. Phys. **229**(4), 978–1016 (2010). https://doi.org/10.1016/j.jcp.2009.09.044
8. Feireisl, E., Klingenberg, C., Markfelder, S.: On the low Mach number limit for the compressible Euler system. SIAM J. Math. Anal. **51**(2), 1496–1513 (2019)
9. Fjordholm, U.S., Mishra, S., Tadmor, E.: Arbitrarily high-order accurate entropy stable essentially nonoscillatory schemes for systems of conservation laws. SIAM J. Numer. Anal. **50**(2), 544–573 (2012)
10. Gresho, P.M., Chan, S.T.: On the theory of semi-implicit projection methods for viscous incompressible flow and its implementation via a finite element method that also introduces a nearly consistent mass matrix. part 2: implementation. Int. J. Numer. Methods Fluids **11**(5), 621–659 (1990). https://doi.org/10.1002/fld.1650110510
11. Guillard, H., Viozat, C.: On the behaviour of upwind schemes in the low Mach number limit. Comput. Fluids **28**(1), 63–86 (1999). https://doi.org/10.1016/S0045-7930(98)00017-6
12. Harten, A.: On the symmetric form of systems of conservation laws with entropy. J. Comput. Phys. **49**, 151–164 (1983)
13. Hughes, T.J., Franca, L., Mallet, M.: A new finite element formulation for computational fluid dynamics: I. symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics. Comput. Methods Appl. Mech. Eng. **54**(2), 223–234 (1986)
14. Ismail, F., Roe, P.L.: Affordable, entropy-consistent Euler flux functions ii: Entropy production at shocks. J. Computat. Phys. **228**(15), 5410–5436 (2009)
15. Jameson, A.: Formulation of kinetic energy preserving conservative schemes for gas dynamics and direct numerical simulation of one-dimensional viscous compressible flow in a shock tube using entropy and kinetic energy preserving schemes. J. Sci. Comput. **34**(2), 188–208 (2008)
16. Kraaijevanger, J.F.B.M.: Contractivity of Runge-Kutta methods. BIT Numer. Math. **31**(3), 482–528 (1991)
17. Li, X., Gu, C.W.: An all-speed Roe-type scheme and its asymptotic analysis of low Mach number behaviour. J. Comput. Phys. **227**(10), 5144–5159 (2008)
18. Miczek, F., Röpke, F.K., Edelmann, P.V.F.: New numerical solver for flows at various Mach numbers. Astron. Astrophys. Rev. **576**, A50 (2015). https://doi.org/10.1051/0004-6361/201425059
19. Oßwald, K., Siegmund, A., Birken, P., Hannemann, V., Meister, A.: L2roe: a low dissipation version of Roe's approximate Riemann solver for low Mach numbers. Int. J. Numer. Methods Fluids (2015). https://doi.org/10.1002/fld.4175.Fld.4175
20. Ray, D., Chandrashekar, P.: Entropy stable schemes for compressible Euler equations. Int. J. Numer. Anal. Model. Ser. B **4**(4), 335–352 (2013)

21. Ray, D., Chandrashekar, P., Fjordholm, U.S., Mishra, S.: Entropy stable scheme on two-dimensional unstructured grids for Euler equations. Commun. Comput. Phys. **19**(5), 1111–1140 (2016)
22. Rieper, F.: A low-Mach number fix for Roe's approximate Riemann solver. J. Comput. Phys. **230**(13), 5263–5287 (2011)
23. Roe, P.L.: Approximate Riemann solvers, parameter vectors, and difference schemes. J. Comput. Phys. **43**(2), 357–372 (1981). https://doi.org/10.1016/0021-9991(81)90128-5
24. Tadmor, E.: The numerical viscosity of entropy stable schemes for systems of conservation laws. i. Math. Comput. **49**(179), 91–103 (1987)
25. Tadmor, E.: Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. Acta Numerica **12**, 451–512 (2003)
26. Toro, E.F.: Riemann Solvers and Numerical Methods for Fluid Dynamics: a Practical Introduction. Springer, Heidelberg (2009)
27. Turkel, E.: Preconditioning techniques in computational fluid dynamics. Ann. Rev. Fluid Mech. **31**, 385–416 (1999). https://doi.org/10.1146/annurev.fluid.31.1.385
28. Zenk, M.: On numerical methods for astrophysical applications. Ph.D. thesis, University of Würzburg (2017). https://opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/16266/file/Zenk_Markus_Dissertation.pdf

Residual Based Method for Sediment Transport

P. Poullet, P. Ramsamy, and M. Ricchiuto

Abstract This contribution deals with a high order Residual Distribution (RD) numerical scheme to simulate sediment transport. The morphodynamic model that has been used, couples shallow–water equations for the fluid flow and the Exner law for the sediment part. Thus, the choice of the approach by a non-conservative hyperbolic system has been made. Different schemes have already been applied to approximate the entropic solution for several test cases [10]. The one proposed in this paper resorts to RD-method, TVD Runge Kutta [27, 31] and stabilisation upwind methods [13], with limiters. It can be viewed as an improvement of the generalized approximate Roe method [8, 14, 29] with some other good properties (Path-conservative, well-balanced...). Numerical results show the ability of the model in 1D to compute accurate solutions and to reproduce some classical test problems. The best results that we obtained, use MinMod flux limiters.

1 Introduction

This work is incorporated within the framework of the study of a sediment transport modelling. One aim of this contribution, is to provide first, an useful simulation tool, in the context of a 1D space-time problem. Considering the geophysical aspect,

P. Poullet (✉) · P. Ramsamy
Université des Antilles, Laboratoire de Mathématiques Informatique et Applications, Campus de Fouillole, 97159 Pointe-à-Pitre Cédex, Guadeloupe
e-mail: pascal.poullet@univ-antilles.fr

P. Ramsamy
e-mail: priscilla.ramsamy@univ-antilles.fr

M. Ricchiuto
Team Cardamom - INRIA, Univ. Bordeaux, CNRS, Bordeaux INP, IMB, UMR 5251, 200 av. de la vieille Tour, 33405 Talence Cédex, France
e-mail: mario.ricchiuto@inria.fr

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*, SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2_9

the sediment transport can be divided into several categories, in this paper we will be interested in bedload transport. The governing equations consist of a system of three equations, modelling the interaction between fluid and sediment in a river. The hydrodynamical component is described by the shallow–water equations (SWE) and the morphodynamical component, is given by a solid transport discharge due to the Exner law with the Grass formula. In this way, the model can be depicted by a non-conservative and non-linear hyperbolic system, and our main objective is to seek numerical solutions in accordance with these specific aspects. In particular, a good scheme for that model, must comply with the well-balanced property and the path-conservative character. Moreover, as the fluid interacts very weakly with the sediment and characteristic velocities being such a different magnitude, long time simulation and high order accuracy (at least second order) are needed.

To treat hyperbolic problem, finite volumes are the most popular methods as the Godunov scheme. For conservation laws, first attempts to propose approximate solver for hyperbolic systems in non-conservative form, were due to Roe [28]. After that, several approaches have been introduced like approximate finite volume Roe with characteristic flux scheme [14], schemes based on exact or incomplete Riemann solvers [5], WENO schemes [32], generalized Roe methods with or without WENO reconstruction [8–10], or kinetic schemes [24].

In parallel to finite volume, another family of methods called Residual Distribution (RD) methods that emerged from Roe's works in the 80s, is used in this paper [11, 28]. Combining advantages from finite volume and finite element methods, their construction allows them nowadays to be monotone, conservative, well-balanced and easily high order accurate [1, 2, 19, 20, 27, 31].

In this contribution, a RD scheme, viewed as a recast of the approximate Riemann solver, called FV-Roe approximation scheme is proposed with simulations. Often used to solve the shallow–water problem, the residual based method is adapted here to solve the coupling problem with sediment transport. And for that the use of a TVD Runge Kutta procedure, but also upwinding and a flux limiter procedures, have been added, to compute weak entropic solution, considering Lax entropy.

To introduce our scheme, the present paper is organized according to the following outline: In Sect. 2, the governing equations are introduced. In Sect. 3, our scheme is proposed. And finally, numerical tests are given, in order to show how accurate the scheme is, but also its well-balanced preserving aspect, for the lake at rest.

2 A Sediment Transport in a Shallow Water

In the context of the study of sediment transport in shallow water, several morphodynamical models can be found in the litterature depending on the way of considering the displacement mode. In the case of bedload transport, among several models [12, 18, 21, 30], the discharge is written by the Grass formula [15] for simplicity. It involves that the critical shear stress is neglected, then the sediment is viewed as starting its own movement as soon as the fluid starts to move. About the hydrody-

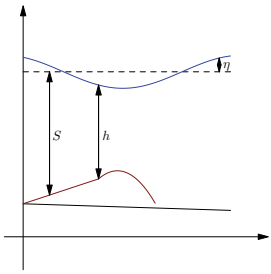


Fig. 1 A sediment layer in a shallow water

namical part, shallow–water equations are considered. The result is an hyperbolic system of three equations.

The governing system for the 1D space-time problem, is as follows

$$\begin{cases} \dfrac{\partial h}{\partial t} + \dfrac{\partial q}{\partial x} = 0 \\ \dfrac{\partial q}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{q^2}{h} + \dfrac{1}{2}gh^2\right) = gh\dfrac{\partial S}{\partial x} - ghS_f \\ \dfrac{\partial S}{\partial t} - \zeta\dfrac{\partial q_b}{\partial x} = 0, \end{cases} \tag{1}$$

where x denote the horizontal variable at the axis of the channel and t the time variable (see [10]). By $h(x,t)$ we denote the height of the water column, $q(x,t)$ is the discharge, g is the gravity constant, S_f models the friction term and ζ a parameter linked to the sediment porosity ($\zeta = 1/(1-\rho_0)$ with ρ_0 the porosity). The third equation of the model describes the sediment transport by the expression of the sediment volume equation, q_b being the solid transport discharge obtained by Grass formula (here, $q_b = A_g(q/h)^3$ with A_g related to the interaction between the fluid and the sediment). The variable $S(x,t)$ is the distance from a given reference level to the bottom layer. A schematic description is provided (see Fig. 1), η denoting the extra height of the water column.

Neglecting S_f in this study, a classical approach is to treat the system (1) as a hyperbolic system with a non-conservative term B:

$$\frac{\partial W}{\partial t} + \mathcal{A}(W)\frac{\partial W}{\partial x} = 0, \tag{2}$$

with $W = W(x,t)$ and $(x,t) \in \mathbf{R} \times \mathbf{R}^+$. In fact, the vector of unknowns is

$$W = (h,\ q,\ S)^T,$$

F is the flux function and, $\mathcal{A}(W)$ equals to the difference between the Jacobian matrix of F and the non-conservative part:

$$\mathcal{A}(W) = \frac{\partial F}{\partial W}(W) - B(W).$$

More precisely,

$$F(W) = \begin{pmatrix} q \\ \frac{q^2}{h} + \frac{1}{2}gh^2 \\ -\zeta q_b \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & gh \\ 0 & 0 & 0 \end{pmatrix}, \quad A(W) = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{q^2}{h^2} + gh & 2\frac{q}{h} & 0 \\ -\zeta \frac{\partial q_b}{\partial h} & -\zeta \frac{\partial q_b}{\partial q} & 0 \end{pmatrix},$$

$$\text{and } \mathcal{A}(W) = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{q^2}{h^2} + gh & 2\frac{q}{h} & -gh \\ -\zeta \frac{\partial q_b}{\partial h} & -\zeta \frac{\partial q_b}{\partial q} & 0 \end{pmatrix}.$$

3 A Residual Based Predictor-Corrector Upwind Discretization for 1D Space-Time Sediment Transport

Following the introduction, the scheme built in this work, can be viewed as an high order recast of the approximate FV-Roe as introduced by Ghidaglia *et al.* in [14]. Instead of using a high order reconstruction, our approach exploits the residual based approach discussed in [26, 27].

To present the scheme, one neglects the friction term, and one considers the hyperbolic system with a non-conservative term. Also, to simplify writing in this section, in integrals the term dx will be omitted (all are space integrals).

One will consider the intervals (computing cells) defined by $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, $i \in \mathbb{Z}$, but also the intervals $I_{i+\frac{1}{2}} = [x_i, x_{i+1}]$. Let the step $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ and that $x_{i-\frac{1}{2}} = \sum_{k=1}^{i-1} \Delta x_k$ is the intercell located at the middle of $I_{i-1} \cup I_i$. Δt is the time step and $t^n = n\Delta t$. As usual, we denote by W_i^n the approximate mean value of W in node x_i and at time t^n. The RD procedure consists of making the computation of the residual, called global, on a cell I_i and then distributing fractions of this quantity to each of its verteces. Under these assumptions, one gets for a linear approximation of $W(x, t^n)$,

$$W_i^n := \frac{1}{\Delta x_i} \int_{I_i} W(x, t^n) \approx W(x_i, t^n).$$

Then, the residual–based predictor corrector method developed in [26, 27] can be written as follows

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x} \left\{ \frac{1}{2}\Theta_i(W^n) + \frac{1}{2}\Theta_i(W^*) \right\} + \Psi_{i-\frac{1}{2}} + \Psi_{i+\frac{1}{2}}, \tag{3}$$

with $\Theta_i(W)$ corresponds to the approximate FV-Roe scheme discrete evolution operator (without the additional term)

$$\Theta_i(W) = \mathcal{P}^+(\mathcal{A}_{i-\frac{1}{2}})\phi_{i-\frac{1}{2}} + \mathcal{P}^-(\mathcal{A}_{i+\frac{1}{2}})\phi_{i+\frac{1}{2}}.$$

Concerning the non-conservative terms, one proceeds by linearization along the path joining W_i and $W_{i\pm1}$, to compute an approximate value of $W_{i\pm\frac{1}{2}} = \frac{1}{2}(W_i + W_{i\pm1})$.

We then denote by $A_{i\pm\frac{1}{2}} = A(W_{i\pm\frac{1}{2}})$, $B_{i\pm\frac{1}{2}} = B(W_{i\pm\frac{1}{2}})$. Also, the fluctuations are:

$$\phi_{i-\frac{1}{2}} = F_i - F_{i-1} - B_{i-\frac{1}{2}}(W_i - W_{i-1}) \text{ and } \phi_{i+\frac{1}{2}} = F_{i+1} - F_i - B_{i+\frac{1}{2}}(W_{i+1} - W_i),$$

and the projectors,

$$\mathcal{P}^{\pm}(\mathcal{A}_{i\mp\frac{1}{2}}) = \frac{1}{2}(I \pm \text{sign}(\mathcal{A}_{i\mp\frac{1}{2}})),$$

with the sign of a matrix computed by eigen-decomposition

$$\text{sign}(\mathcal{A}_{i\mp\frac{1}{2}}) = \mathcal{K}_{i\mp\frac{1}{2}}\text{sign}(\mathcal{L}_{i\mp\frac{1}{2}})\mathcal{K}^{-1}_{i\mp\frac{1}{2}}.$$

The matrix $\mathcal{K}$ gathering the eigenvectors of the Roe matrix (along the path), sign($\mathcal{L}$) is the diagonal matrix whose coefficients are the sign of the eigenvalues (see [23] for example). We denote by W^*, a predicted value of the solution that has been obtained, from the upwind scheme,

$$W_i^* = W_i^n - \frac{\Delta t}{\Delta x_i}\Theta_i(W^n).$$

Actually, if we introduce a parallel approach using Galerkin method by seeking a piecewise linear solution, in a domain $\Omega =]0, L[$,

$$W_h(x,t) = \sum_i \varphi_i(x) W_i(x_i, t), \qquad \text{and} \quad F_h = F(W_h),$$

with φ_i representing the standard Lagrange basis fonctions associated to the node x_i. One replaces the unknown solution by its approximate solution by finite element from the variational form

$$\int_\Omega \varphi_i \partial_t W_h + \int_\Omega \varphi_i \partial_x F_h - \int_\Omega \varphi_i B(W_h)\partial_x W_h = 0. \tag{4}$$

And one can notice that by linear approximation,

$$\int_\Omega \varphi_i \partial_x F_h = \int_{I_{i-\frac{1}{2}}} \varphi_i \partial_x F_h + \int_{I_{i+\frac{1}{2}}} \varphi_i \partial_x F_h \approx \frac{F_i - F_{i-1}}{2} + \frac{F_{i+1} - F_i}{2}. \quad (5)$$

Hence, the resulting non-stabilized method reads

$$\int_\Omega \varphi_i \partial_t W_h + \frac{1}{2}\phi_{i+\frac{1}{2}} + \frac{1}{2}\phi_{i-\frac{1}{2}} = 0, \quad (6)$$

with the approximation of the first term $\Delta x_i \frac{dW_i}{dt}$, one recovers the mass lumping process.

But as it is known that Galerkin method suffers from lack of stability, one adds a residual based stabilization in the spirit of the streamline upwind method or Streamline Upwind Petrov Galerkin (SUPG) [17, 26, 27].

For a node x_i, the stabilization operator $\mathcal{S}_i$ reads

$$\mathcal{S}_i = \int_\Omega \mathcal{A}(W_h)\partial_x \varphi_i \mathcal{T} \tilde{r} \quad \text{with} \quad \tilde{r} = \partial_t W_h + \partial_x F_h - B(W_h)\partial_x W_h,$$

the matrix $\mathcal{T}$ being a scaling factor guaranteeing the uniform boundedness of the stabilization w.r.t. the residual. As before, explicit computable expressions are obtained when introducing the linear finite element approximation and introducing appropriate mean value linearizations of the matrices that appear. Recalling that for a linear approximation $\partial_x \varphi_i\big|_{I_{i\pm\frac{1}{2}}} = \mp 1/\Delta x_{i\pm\frac{1}{2}}$, the stabilization term can be evaluated as

$$\mathcal{S}_i = \mathcal{A}_{i+\frac{1}{2}}\mathcal{T}_{i+\frac{1}{2}}\left(\int_{I_{i+\frac{1}{2}}} \partial_t W_h + \phi_{i+\frac{1}{2}}\right) - \mathcal{A}_{i-\frac{1}{2}}\mathcal{T}_{i-\frac{1}{2}}\left(\int_{I_{i-\frac{1}{2}}} \partial_t W_h + \phi_{i-\frac{1}{2}}\right) \quad (7)$$

In one dimension, a typical definition of the scaling matrix $\mathcal{T}$, also used here, being the following

$$\mathcal{T}_{i\pm\frac{1}{2}} = \frac{1}{2}\Delta x_{i\pm\frac{1}{2}} \operatorname{sign}(\mathcal{A}_{i\pm\frac{1}{2}})\mathcal{A}^{-1}_{i\pm\frac{1}{2}} = \frac{\Delta x_{i\pm\frac{1}{2}}}{2}|\mathcal{A}_{i\pm\frac{1}{2}}|^{-1}, \quad (8)$$

the complete semi-discrete (in space) equations read

$$\int_{I_{i-\frac{1}{2}}} \left(\varphi_i + \operatorname{sign}(\mathcal{A}_{i-\frac{1}{2}})\right)\partial_t W_h + \int_{I_{i+\frac{1}{2}}} \left(\varphi_i - \operatorname{sign}(\mathcal{A}_{i+\frac{1}{2}})\right)\partial_t W_h = -\left(P^+(\mathcal{A}_{i-\frac{1}{2}})\phi_{i-\frac{1}{2}} + P^-(\mathcal{A}_{i+\frac{1}{2}})\phi_{i+\frac{1}{2}}\right). \quad (9)$$

As it has been explained in [27], a simplified stabilization step can be added in the Galerkin process with the Runge-Kutta scheme without any loss of accuracy (here, an explicit RK2 of second order accuracy is used). Thus, the residual

$$r^{n+1} := \frac{W_h^{n+1} - W_h^n}{\Delta t} + \frac{1}{2}(\partial_x F_h - B_h \partial_x W_h)^* + \frac{1}{2}(\partial_x F_h - B_h \partial_x W_h)^n \tag{10}$$

can be replaced with a simplified residual

$$r^* := \frac{W_h^* - W_h^n}{\Delta t} + \frac{1}{2}(\partial_x F_h - B_h \partial_x W_h)^* + \frac{1}{2}(\partial_x F_h - B_h \partial_x W_h)^n, \tag{11}$$

to design the scheme by computing

$$\int_\Omega \varphi_i r^{n+1} + \int_{I_{i-\frac{1}{2}} \cup I_{i+\frac{1}{2}}} A(W_h)\partial_x \varphi_i \, \mathcal{T} \, r^* = 0. \tag{12}$$

Using an explicit scheme by performing mass lumping in the predictor step (6), and with midpoint rule to evaluate the integrals, one obtains after few calculations and recast Eq. (3), that one recalls

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\left\{\frac{1}{2}\Theta_i(W^n) + \frac{1}{2}\Theta_i(W^*)\right\} + \Psi_{i-\frac{1}{2}} + \Psi_{i+\frac{1}{2}},$$

with the aim to define $\Psi_{i\pm\frac{1}{2}}$ as:

$$\begin{aligned}\psi_{i\pm\frac{1}{2}} =& \frac{\Delta t}{\Delta x}\int_{I_{i\pm\frac{1}{2}}} \left(\varphi_i - \frac{1}{2}(I \mp \text{sign}(\mathcal{A}_{i\pm\frac{1}{2}}))\right)\frac{W_h^* - W_h^n}{\Delta t} \\ \approx& \frac{1}{2}\left(W_i^* - W_i^n - (I \mp \text{sign}(\mathcal{A}_{i\pm\frac{1}{2}}))(W_{i\pm\frac{1}{2}}^* - W_{i\pm\frac{1}{2}}^n)\right).\end{aligned}$$

However, comparing to the FV-Roe scheme, the additional terms $\Psi_{i\pm\frac{1}{2}}$ that derive from residual stabilization are of second order, that cannot always match with non regular solution. To avoid their effects accross discontinuous features, cell based limiters have been introduced and finally,

$$\Psi_{i\pm\frac{1}{2}} = \delta_{i\pm\frac{1}{2}}\psi_{i\pm\frac{1}{2}}$$

with $\delta_{i\pm\frac{1}{2}}$ computed by means of a standard finite volume limiters using values of the MUSCL MinMod flux limiter function [7].

4 Numerical Results

We present the numerical results of several reference problems. In this section, we aim at validating our numerical scheme and highlighting its characteristic properties against classical tests. The first test consists in proving an approximate well-balanced property of our scheme. We then underline the ability of our scheme to simulate a parabolic sediment transport until a discontinuous solution is obtained. We also prove its high order accuracy by means of an order test problem that has been discussed in [10]. Finally, we prove that our scheme is capable of faithfully reproducing a dam-break problem over a wet bottom topography [4].

4.1 Test of Well-Balanced Property

To check the property, the following numerical test is used [25]. It deals with the ability of the scheme to reproduce the behaviour of the steady state. Thus, if the numerical scheme is well-balanced, a small difference should be observed between the initial solution and the solution obtained at the final instant.

For this, the interval [0, 10] is assumed as the physical domain and the simulation is performed up to $T = 0.5\,\mathrm{s}$ with 100 and 200 cells. A discontinuity in the bed is assumed as the initial condition, so the thickness of the sediment layer is considered,

$$z_b(x,0) = \begin{cases} 4 & \text{if} \quad 4 \leq x \leq 8 \\ 0 & \text{elsewhere} \end{cases} \tag{13}$$

and

$$q(x,0) = 0, \quad h(x,0) + z_b(x,0) = 10.$$

The results in Table 1, show that the scheme preserves the approximate well-balanced property [22]. The differences between the initial solution and the solution at final time are very small. More precisely, the ratio is of around 1.5 between the doubled gridpoints and the coarser one. The accuracy seems to be of order 1.5 (Table 1).

Table 1 Accuracy of the scheme for the well-balanced test property

Precision	L^2-error h	Ratio	L^2-error q	Ratio
100	8.6052×10^{-16}	-	7.1712×10^{-15}	-
200	5.2050×10^{-16}	1.65	5.8743×10^{-15}	1.22

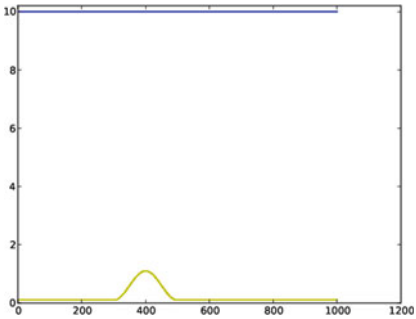


Fig. 2 The dune at initial time

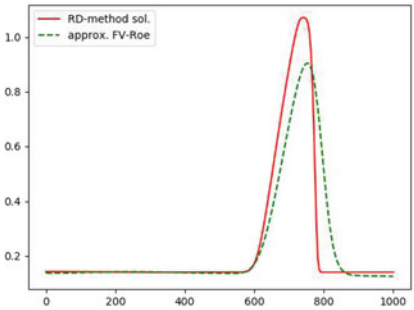


Fig. 3 z_b at 700 s for RD scheme and FV-Roe

4.2 A 1D Space-Time Dune Test Case

To verify the shock capturing property, the classical transport of parabolic sediment layer has been taken. For this case proposed in [7, 10, 16], the interval [0, 1000] is assumed as the physical domain and a strong interaction between the fluid and the sediment is taken ($A_g = 1$). The initial conditions are given as follows (see Fig. 2): the bathymetry is of 10 m, $q(x,t) = 10$ for the discharge of the fluid, $h(x,t) = 10 - z_b(x,t)$ for the water column height with the sediment layer thickness ,

$$z_b(x,0) = \begin{cases} 0.1 + \sin^2\left(\frac{\pi(x-300)}{200}\right) & \text{if} \quad 300 \leq x \leq 500 \\ 0.1 & \text{elsewhere} \end{cases}. \tag{14}$$

Numerical solutions are generated up to $T = 700$ s. As shown in Fig. 3, the solution of our scheme (solid line) is compared to the solution of the FV-Roe approximation (dotted line), and we note that the RD scheme is less diffusive than the first and the shock is more visible (thanks to the limiters). Therefore, the results show that our numerical scheme seems to satisfy the shock capturing property for this test. It can be noted that it is also stable, more precisely the RD scheme does not generate oscillations even though the interaction between the fluid and the sediment is strong (somewhat artificial with the initial conditions chosen [16]). The comparison between the evolution of the dune between the RD scheme and the SRNH scheme (a second-

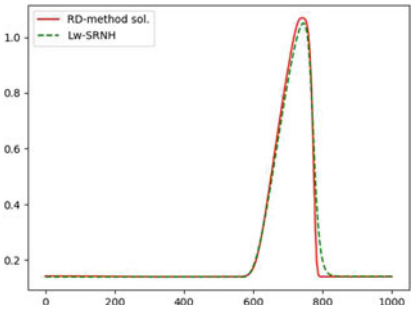


Fig. 4 z_b at 700 s for RD scheme and SRNH

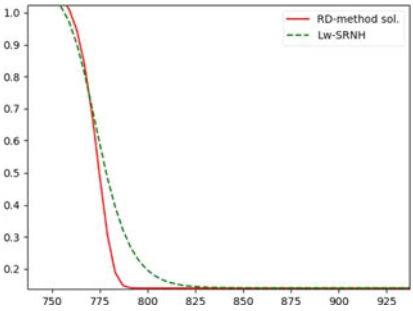


Fig. 5 z_b at 700 s for RD scheme and SRNH: zoom on the shock

order scheme extracted from [6], see Figs. 4 and 5) seems to confirm that our scheme is of the second order and is less diffusive than the other.

4.3 A Test of Order

To check the accuracy of the numerical scheme, let us introduce the following one-dimensional problem for which initial conditions are:

$$q(0,x) = 0, \qquad h(0,x) = 2 - 0.1\exp(-x^2), \qquad z_b(0,x) = 0.1 - 0.01\exp(-x^2). \tag{15}$$

This test problem has been considered in a previous work by Castro Diaz *et al.* [10], and as the exact solution is unknown, we use a reference solution obtained by a fine mesh of 5120 volumes (as it has been done in their work) with a medium interaction ($A_g = 0.3$).

Numerically, the results for h, q and the sediment layer thickness (z_b) and different error norms, show that the second order of accuracy seems to be obtained (see Tables 2, 3 and 4).

Table 2 Accuracy of the RD-scheme with distributed residual scheme for h

# gridpts	L^1-error	Order	L^2-error	Order	L^∞-error	Order
20	5.628×10^{-3}	-	8.600×10^{-3}	-	2.088×10^{-2}	-
40	2.421×10^{-3}	1.22	4.159×10^{-3}	1.05	1.086×10^{-2}	0.94
80	7.918×10^{-4}	1.61	1.408×10^{-3}	1.56	4.036×10^{-3}	1.43
160	2.072×10^{-4}	1.93	3.815×10^{-4}	1.88	1.211×10^{-3}	1.74
320	5.257×10^{-5}	1.98	9.693×10^{-5}	1.98	3.159×10^{-4}	1.94
640	1.314×10^{-5}	2.00	2.423×10^{-5}	2.00	7.927×10^{-5}	1.99

Table 3 Accuracy of the RD-scheme with distributed residual scheme for q, the discharge

# gridpts	L^1-error	Order	L^2-error	Order	L^∞-error	Order
20	2.350×10^{-2}	-	3.771×10^{-2}	-	9.248×10^{-2}	-
40	1.042×10^{-2}	1.17	1.835×10^{-2}	1.04	4.783×10^{-2}	0.95
80	3.435×10^{-3}	1.60	6.205×10^{-3}	1.56	1.783×10^{-2}	1.42
160	8.995×10^{-4}	1.93	1.681×10^{-3}	1.88	5.349×10^{-3}	1.73
320	2.281×10^{-4}	1.98	4.269×10^{-4}	1.98	1.397×10^{-3}	1.94
640	5.697×10^{-5}	2.00	1.067×10^{-4}	2.00	3.510×10^{-4}	1.99

Table 4 Accuracy of the RD-scheme with distributed residual scheme for z_b, the height of the sediment layer

# gridpts	L^1-error	Order	L^2-error	Order	L^∞-error	Order
20	2.969×10^{-5}	-	5.443×10^{-5}	-	1.423×10^{-4}	-
40	1.265×10^{-5}	1.23	2.739×10^{-5}	0.99	8.168×10^{-5}	0.80
80	4.469×10^{-6}	1.50	9.014×10^{-6}	1.60	2.861×10^{-5}	1.51
160	1.200×10^{-6}	1.90	2.536×10^{-6}	1.83	8.578×10^{-6}	1.74
320	3.019×10^{-7}	1.99	6.409×10^{-7}	1.98	2.217×10^{-6}	1.95
640	7.516×10^{-8}	2.01	1.596×10^{-7}	2.00	5.571×10^{-7}	1.99

4.4 A Dam Break Test over a Wet Bottom Topography

In this classical test case ([4]) a dam break is considered over a flat wet bottom, in a channel of 10m long. A low interaction between the fluid and the sediment is taken ($A_g = 0.005$), and the initial conditions are,

$$h(x,0) = \begin{cases} 2 & \text{if } x \leq 5m \\ 0.125 & \text{if } x > 5m \end{cases}, \tag{16}$$

$q(x,0) = 0m/s$ and the bottom topography $z_b = 0m$. The numerical test is performed until $T = 1s$. The results confirm that our scheme keep the stability as attempted for the coupled approaches (see Fig. 6, 7), in comparison the approach by splitting [3]. The accuracy of our RD scheme is of course better than those obtained by the FV-Roe scheme (8). More precisely, as expected and considering

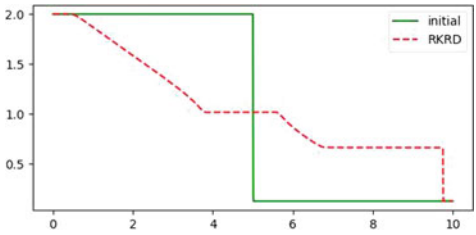


Fig. 6 Dam break over a wet bottom: h at $t = 1s$ for RD scheme with 1000 cells

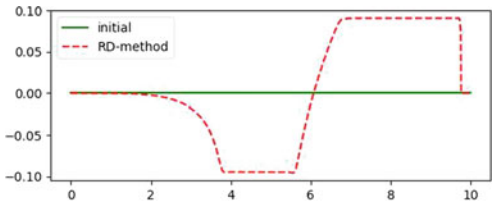


Fig. 7 Dam break over a wet bottom: z_b at $t = 1s$ for RD scheme with 1000 cells

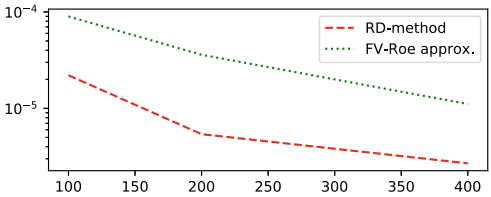


Fig. 8 Dam break over a wet bottom: error analysis for RD scheme

the bottom firstly (Fig. 6) the solution computed with the RD scheme forms two plates without any oscillation (in a critical region includes the interval $[4, 6]$ with $-0.1 < z_b < -0.05$, and in a region where $x > 6.5$ with $z_b > 0.5$). Then, for the free surface (Fig. 7), the solution of the RD scheme (dashed line) is decreasing along the time, and a plate is reached without oscillation in the critical region $[4, 6]$. For both unknowns, a comparison with the FV-Roe approximation (dotted line) is proposed, underlying that the results are quite similar. However the RD scheme is more accurate for the bottom (see Fig. 8 for which the L^2 errors are produced from a referent solution computed with 2000 elements grid).

5 Concluding Remarks

This contribution proposed a new predictor-corrector scheme, based on the residual, to simulate a sediment transport problem. Numerical tests have highlighted its high order accuracy, its approximate well-balancedness property and its stability for some test problems. In particular, the solutions obtained for the problem of dam-break with wet topography are sharp. Work is in progress to take into account dry bottoms, for example. The extension of this model to take into account a coastal configuration (2D physical domain), by parallel programming, will also be done in a future contribution.

Acknowledgements Computational tests have been performed using Wahoo, the server of the Centre Commun de Calcul Intensif of the Université des Antilles.

References

1. Abgrall, R., Ricchiuto, M.: High-Order Methods for CFD, pp. 1–54. John Wiley, Ltd. (2017). https://doi.org/10.1002/9781119176817.ecm2112
2. Arpaia, L., Ricchiuto, M., Abgrall, R.: An ALE formulation for explicit Runge-Kutta residual distribution. J. Sci. Comput. **32**, 502–547 (2015)
3. Audusse, E., Berthon, C., Chalons, C., et al.: Sediment transport modelling: relaxation schemes for Saint-Venant–Exner and three layer models. In: CEMRACS2011: Multiscale Coupling of Complex Models in Scientific Computing, *ESAIM: Proc*, vol. 38, pp. 78–98. EDP Sciences (2012)
4. Audusse, E., Chalons, C., Ung, P.: A simple three-wave approximate Riemann solver for the Saint-Venant-Exner equations. Int. J. Numer. Method Fluids **87**, 508–528 (2018)
5. Benkhaldoun, F., Sahmim, S., Seaid, M.: Solution on the sediment transport equations using a finite volume method based on sign matrix. SIAM J. Sci. Comput. **31**(4), 2866–2889 (2009)
6. Benkhaldoun, F., Sari, S., Seaid, M.: A flux-limiter method for dam-break flows over erodible sediment beds. Appl. Math. Model. **36**(10), 4847–4861 (2012)
7. Bilanceri, M., Beux, F., Elmahi, I., et al.: Linearized implicit time advancing and defect correction applied to sediment transport simulations. Comput. Fluids **63**, 82–104 (2012)
8. Castro, M., Gallardo, J., Parés, C.: High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative products. Applications to shallow water systems. Math. Comp. **75**, 1103–1134 (2006)
9. Castro, M., Pardo, A., Parés, C., Toro, E.: On some fast well-balanced first order solvers for nonconservative systems. Math. Comput. **79**(271), 1427–1472 (2010)
10. Castro Diaz, M., Fernandez-Nieto, E., Ferreiro, A.: Sediment transport models in shallow water equations and numerical approach by high order finite volume methods. Comput. Fluids **37**, 299–316 (2008)
11. Deconinck, H., Ricchiuto, M.: Residual distribution schemes: foundations and analysis, chapter 19. John Wiley & Sons, Ltd. (2007)
12. Einstein, H.: Formulas for the transportation of bed load. Trans. ASCE **107**, 561–575 (1942)
13. Fillipini, A., Ricchiuto, M.: Upwind residual discretization of enhanced Boussinesq equations for wave propagation over complex bathymetries. J. Comput. Phys. **271**(15), 306–341 (2014)
14. Ghidaglia, J.M., Kumbaro, A., Le Coq, G.: Une méthode de volumes finis à flux caractéristiques pour la résolution numérique des systèmes hyperboliques de lois de conservation. C.R. Acad. Sc. Paris **322**(I), 981–988 (1996)
15. Grass, A.: Sediments transport by waves and currents. SERC London Cent. Mar, Technology (1981)
16. Hudson, J.: Numerical techniques for morphodynamic modelling. Ph.D. thesis, University of Reading (2001)
17. Hughes, T., Franca, L., Mallet, M.: Finite element formulation for computational fluid dynamics: I symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics. Comput. Methods Appl. Mech. Eng. **54**, 223–234 (1986)
18. Kalinske, A.: Movement of sediment as bed load in rivers. Trans. AGU **28**(4), 615–620 (1947)
19. Lin, J., Abgrall, R., Qiu, J.: High order residual distribution for steady state problems for hyperbolic conservation laws. J. Sci. Comput. **79**, 891–913 (2019)
20. Mazaheri, A., Nishikawa, H.: Improved second-order hyperbolic residual-distribution scheme and its extension to third-order on arbitrary triangular grids. J. Comput. Phys. **300**, 455–491 (2015)

21. Meyer-Peter, E., Müller, R.: Formulas for Bed-Load transport. In: Report on 2nd meeting on international association on hydraulic structures research. Stockholm. IAHR (1948)
22. Munoz-Ruiz, M., Parés, C.: On the convergence and well-balanced property of path-conservative numerical schemes for systems of balance laws. J. Sci. Comput. **48**, 274–295 (2011)
23. Parés, C., Castro, M.: On the well-balance property of Roe's method for nonconservative hyperbolic systems. applications to shallow–water systems. ESAIM: M2AN **38**(5), 821–852 (2004)
24. Perthame, B., Simeoni, C.: A kinetic scheme for the Saint-Venant system with a source term. Calcolo **38**(4), 201–231 (2001)
25. Qian, S., Li, G., Shao, F., Niu, Q.: Well-balanced central WENO schemes for the sediment transport model in shallow water. Comput. Geosci. **22**, 763–773 (2018)
26. Ricchiuto, M.: An explicit residual based approach for shallow water flows. J. Comput. Phys. **280**, 306–344 (2015)
27. Ricchiuto, M., Abgrall, R.: Explicit Runge-Kutta residual distribution schemes for time dependent problems: second order case. J. Comput. Phys. **229**, 5653–5691 (2010)
28. Roe, P.: Approximate Riemann solvers, parameter vectors and difference schemes. J. Comput. Phys. **43**, 357–372 (1981)
29. Toumi, I.: A weak formulation of Roe's approximate Riemann solver. J. Comput. Phys. **102**(2), 360–373 (1992)
30. Van Rijn, L.: Sediment transport (i): bed load transport. J. Hydraul. Eng. **110**, 1431–1456 (1984)
31. Warzynski, A., Hubbard, M., Ricchiuto, M.: Runge-Kutta residual distribution schemes. J. Sci. Comput. **62**, 772–802 (2015)
32. Xing, Y., Shu, C.: High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms. J. Comput. Phys. **214**, 567–598 (2006)

New Flow Models

Pseudo-compressibility, Dispersive Model and Acoustic Waves in Shallow Water Flows

Anne-Sophie Bonnet-Ben Dhia, Marie-Odile Bristeau, Edwige Godlewski, Sébastien Impériale, Anne Mangeney, and Jacques Sainte-Marie

Abstract In this paper we study a dispersive shallow water type model derived from the free surface compressible Navier-Stokes system. The compressible effects allow to capture the acoustic-like waves propagation and can be seen as a relaxation of an underlying incompressible model. The first interest of such a model is thus to capture both acoustic and water waves. The second interest lies in its numerical approximation. Indeed, at the discrete level, the pseudo-compressibility terms circumvent the resolution of an elliptic equation to capture the non-hydrostatic part of the pressure. This drastically reduces the cost of the numerical resolution of dispersive models especially in 2d and 3d.

A.-S. Bonnet-Ben Dhia · S. Impériale
Inria Saclay, rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France
e-mail: Anne-Sophie.Bonnet-Bendhia@ensta-paris.fr

S. Impériale
e-mail: Sebastien.Imperiale@inria.fr

M.-O. Bristeau · E. Godlewski · A. Mangeney · J. Sainte-Marie (✉)
Inria Paris, 2 rue Simone Iff, 75589 Paris Cedex 12, France
e-mail: Jacques.Sainte-Marie@inria.fr

M.-O. Bristeau
e-mail: Marie-Odile.Bristeau@inria.fr

E. Godlewski
e-mail: Edwige.Godlewski@upmc.fr

A. Mangeney
e-mail: Anne.Mangeney@ipgp.fr

Sorbonne University, Paris-Diderot University, CNRS, Laboratoire Jacques-Louis Lions, 75005 Paris, France

S. Impériale
Laboratoire de mécanique des solides, Route de Saclay, 91120 Palaiseau, France

A. Mangeney
Institut de Physique du Globe de Paris, Seismology Group, Paris Diderot University, 1 rue Jussieu, 75005 Paris, France

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*,
SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2_10

Keywords Shallow water flows · Dispersive models · Compressible models · Water and acoustic waves · Projection-correction schemes · Finite volumes

Math. classification. 65M12 · 74S10 · 76M12 · 35L65 · 35Q30 · 35Q35 · 76D05 · 76Q05

1 Presentation

The non linear shallow water model with topography [7] is widely used to describe geophysical flows and an extensive literature exists for its numerical approximation [3, 6, 10, 22, 25]. But the classical shallow water equations rely on the hydrostatic assumption and many shallow water type models taking into consideration the dispersive effects have been proposed and studied in the literature, see [2, 8, 9, 12, 13, 19, 23, 27, 28], the list being non-exhaustive.

Considering a two-dimensional domain $\Omega \subset \mathbb{R}^2$ delimited by the boundary $\Gamma = \Gamma_{in} \cup \Gamma_{out} \cup \Gamma_s$ as described in Fig. 1-*(a)*, some of the authors have proposed a family of 2d shallow water dispersive models written under the form [2]

$$\frac{\partial h}{\partial t} + \frac{\partial (hu)}{\partial x} + \frac{\partial (hv)}{\partial y} = 0, \tag{1}$$

$$\frac{\partial (hu)}{\partial t} + \frac{\partial}{\partial x}\left(hu^2 + \frac{g}{2}h^2 + hp\right) + \frac{\partial (huv)}{\partial y} = -(gh + \frac{\gamma^2}{2}p)\frac{\partial z_b}{\partial x}, \tag{2}$$

$$\frac{\partial (hv)}{\partial t} + \frac{\partial (huv)}{\partial x} + \frac{\partial}{\partial y}\left(hv^2 + \frac{g}{2}h^2 + hp\right) = -(gh + \frac{\gamma^2}{2}p)\frac{\partial z_b}{\partial y}, \tag{3}$$

$$\frac{\partial (hw)}{\partial t} + \frac{\partial (huw)}{\partial x} + \frac{\partial (hvw)}{\partial y} = \gamma p, \tag{4}$$

$$\gamma w = -h\frac{\partial u}{\partial x} + \frac{\gamma^2}{2}u\frac{\partial z_b}{\partial x} - h\frac{\partial v}{\partial y} + \frac{\gamma^2}{2}v\frac{\partial z_b}{\partial y}, \tag{5}$$

where $\mathbf{u}(t, \mathbf{x}) = (u, v, w)^T$ is the velocity of the fluid with $\mathbf{x} = (x, y)$, p is the non-hydrostatic part of the fluid pressure, the total pressure is given by $p_{tot} = gh/2 + p$ and g represents the gravity acceleration. The value of the parameter $\gamma \in \mathbb{R}$ will be discussed in Remark 1. The water depth (resp. the topography profile) is denoted $h(t, \mathbf{x})$ (resp. $z_b(\mathbf{x})$) and the free surface is defined by (see Fig. 1-*(b)*)

$$\eta(t, \mathbf{x}) := h(t, \mathbf{x}) + z_b(\mathbf{x}). \tag{6}$$

For smooth solutions, the system (1)–(5) satisfies the following energy balance

$$\frac{\partial E}{\partial t} + \nabla_0 \cdot \left(\mathbf{u}(E + \frac{g}{2}h^2 + hp)\right) = 0, \tag{7}$$

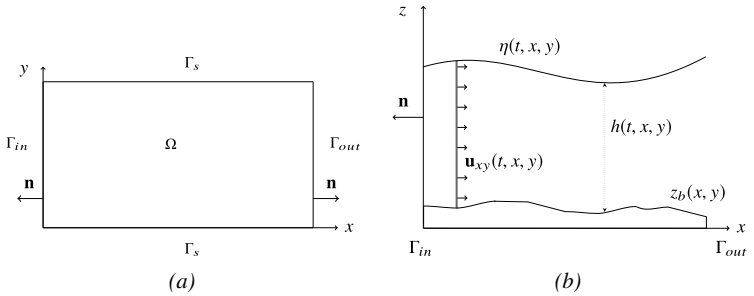


Fig. 1 Model domain and notations, **a** view from above and **b** vertical cross section

with the operator $\nabla_0 = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, 0)^T$ and

$$E = h(u^2 + v^2 + w^2)/2 + g(\eta^2 - z_b^2)/2. \tag{8}$$

The system (1)–(5) defines a family $\{\mathcal{M}_\gamma\}$ of dispersive models written in the more compact form

$$\frac{\partial h}{\partial t} + \nabla_0 \cdot (h\mathbf{u}) = 0, \tag{9}$$

$$\frac{\partial (h\mathbf{u})}{\partial t} + \nabla_0 \cdot (h\mathbf{u} \otimes \mathbf{u}) + \nabla_0(\frac{g}{2}h^2) + \nabla_{sw}^\gamma p = -gh\nabla_0 z_b, \tag{10}$$

$$\mathrm{div}_{sw}^\gamma(\mathbf{u}) = 0, \tag{11}$$

where the shallow water versions of the gradient and divergence operators are defined by

$$\nabla_{sw}^\gamma f = \begin{pmatrix} h\frac{\partial f}{\partial x} + f\frac{\partial \zeta}{\partial x} \\ h\frac{\partial f}{\partial y} + f\frac{\partial \zeta}{\partial y} \\ -\gamma f \end{pmatrix}, \tag{12}$$

$$\mathrm{div}_{sw}^\gamma(\mathbf{w}) = \frac{\partial (hw_1)}{\partial x} + \frac{\partial (hw_2)}{\partial y} - w_1\frac{\partial \zeta}{\partial x} - w_2\frac{\partial \zeta}{\partial y} + \gamma w_3, \tag{13}$$

for $\mathbf{w} = (w_1, w_2, w_3)^T$ and

$$\zeta = h + \frac{\gamma^2}{2} z_b. \tag{14}$$

Whereas ζ depends on γ, for the sake of simplicity, we have adopted a simplified notation and ζ_γ is replaced by ζ.

The model studied in this paper consists in a compressible version of the model (9)–(11) where the divergence free constraint (11) is replaced by an evolution

equation—a relaxed version of (11)—modeling the propagation of acoustic-type waves.

Remark 1 The value of the parameter γ is discussed in [2]. Here we just recall the two extreme hydraulic regimes that can be represented by shallow water models. First the case where $u \ll \sqrt{gh}$ i.e. the fluid velocity is very small compared to the water wave velocity or equivalently the Froude number is very low. In this situation the value $\gamma = \sqrt{3}$ is well adapted since $\mathcal{M}_{\sqrt{3}}$ corresponds to the well known Green-Naghdi model [23]. Another typical situation is the case of advection dominated flows—u cannot be neglected with respect to $\sqrt{gh}$—where the value $\gamma = 2$ is more appropriate.

The numerical analysis of the system (9)–(11) is studied in [2] and a numerical scheme based on a projection-correction scheme [14] has been proposed. Since the model (9)–(11) appears as an extension of the classical Saint-Venant system, the hyperbolic part is treated using a finite volume approach—explicit in time—coupled with the resolution of a saddle point problem—implicit in time—corresponding to an elliptic-type equation for the contribution of the dispersive terms.

Because of the divergence free constraint (11) used to approximate the non-hydrostatic part of the pressure p, an implicit treatment is natural (see Sect. 3.2) but it significantly increases the computational costs. Indeed, an explicit in time scheme constrained by a CFL condition is required for the approximation of the hyperbolic part implying small time steps but simple computations of the numerical fluxes. Whereas the dispersive terms are obtained though the resolution of an elliptic equation for the whole domain. Therefore, for the numerical approximation of the model (9)–(11) over a 2d geometrical domain discretized with N cells, at each time step we have to compute $O(N)$ numerical fluxes and to perform the resolution of a linear symmetric problem. For a stationary linear symmetric problem having at our disposal a good preconditioner, the resolution cost can be estimated as $O(N \log N)$ computations but in our situation, the matrices depend on time—and hence have to be built at each time step—and we do not have any high-performance preconditioner. Hence the computational costs can be estimated as $O(N^{3/2})$, the resolution of the elliptic part becoming very limitative.

In this paper we propose, starting from the compressible Navier-Stokes equations, a modified version of (9)–(11) allowing to propagate both water and acoustic-type waves. The proposed model consists in modifying Eq. (11) in order to include compressibility effects. The new formulation has another advantage since it is possible to discretize it with a fully explicit time scheme and the computational costs are asymptotically $O(N/\sqrt{\varepsilon})$, $\sqrt{\varepsilon}$ being a parameter that will be precised later. Even if the parameter ε can be small, in 2d cases or with fine meshes we have $\varepsilon N \gg 1$ and hence $O(N/\sqrt{\varepsilon}) \ll O(N^{3/2})$.

This paper is organized as follows. First starting from the 3d compressible Navier-Stokes equations, we derive a 2d shallow water model where the acoustic waves—that can be seen as pseudo-compressibility effects—are considered. Then a numerical scheme—explicit in time—is proposed for this 2d model and its properties are

studied. Some stability properties—especially a discrete entropy equality—for the proposed scheme are established in the 1d context. Finally for a well known test case, an illustration comparing the implicit strategy and the resolution of the pseudo-compressible model are presented and the associated computational costs are given.

2 A Compressible and Dispersive Model in Shallow Water Context

In this section, we derive a shallow water approximation of the 3d compressible Navier-Stokes with free surface. The model obtained in Proposition 5 propagates both water and acoustic waves and its dispersive properties are studied. Finally, considering the acoustic velocity is very large compared to the gravity wave velocity, we propose a new formulation as a pseudo-compressible shallow water dispersive model.

2.1 The Compressible Navier-Stokes-Fourier System

We consider the classical compressible Navier-Stokes system describing a free surface gravitational 3d flow over a bottom topography $z_b(x, y)$ (see Fig. 2),

$$\frac{\partial \tilde{\rho}}{\partial t} + \nabla \cdot (\tilde{\rho}\mathbf{U}) = 0, \tag{15}$$

$$\frac{\partial (\tilde{\rho}\mathbf{U})}{\partial t} + \nabla \cdot (\tilde{\rho}\mathbf{U} \otimes \mathbf{U}) + \nabla \tilde{p} - \nabla \cdot \sigma = \tilde{\rho}\mathbf{g}, \tag{16}$$

$$\frac{\partial}{\partial t}\left(\tilde{\rho}\frac{|\mathbf{U}|^2}{2} + \tilde{\rho}\tilde{e}\right) + \nabla \cdot \left(\left(\tilde{\rho}\frac{|\mathbf{U}|^2}{2} + \tilde{\rho}\tilde{e} + \tilde{p} - \sigma\right)\mathbf{U}\right) = -\nabla \cdot Q_{\tilde{T}} + \tilde{\rho}\mathbf{g} \cdot \mathbf{U}, \tag{17}$$

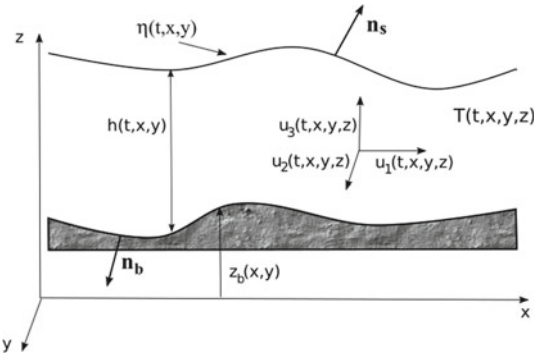


Fig. 2 Flow domain with water height $h(t, \mathbf{x})$, free surface $\eta(t, \mathbf{x})$ and bottom $z_b(\mathbf{x})$

where $\mathbf{U} = (u_1, u_2, u_3)^T$ is the velocity, $\tilde{\rho}$ is the density, $\tilde{p}$ is the fluid pressure, σ is the viscosity stress and $\mathbf{g} = (0, 0, -g)^T$ represents the gravity forces. The internal specific energy is denoted by $\tilde{e}$, the temperature by $\tilde{T}$. The symbol ∇ denotes $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)^T$. In the following, we will also use the notation $\mathbf{v} = (u_1, u_2)^T$ for the horizontal velocity and $\nabla_{x,y}$ corresponds to the projection of ∇ on the horizontal plane i.e. $\nabla_{x,y} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)^T$. The square norm of the velocity vector is $|\mathbf{U}|^2 = u_1^2 + u_2^2 + u_3^2$.

The term $\tilde{\rho}\mathbf{g} \cdot \mathbf{U} = -\tilde{\rho} g u_3$ in (17) prevents this equation from being directly a local energy conservation law. But multiplying the mass conservation (15) by z we get the identity

$$\frac{\partial(z\tilde{\rho})}{\partial t} + \nabla \cdot (z\tilde{\rho}\mathbf{U}) = \tilde{\rho} u_3. \tag{18}$$

Computing the integral along the vertical axis of relation (18) and using the boundary conditions (24), (22)—that are described below—one obtains

$$\frac{\partial}{\partial t} \int_{z_b}^{\eta} g z \tilde{\rho} dz + \nabla_{x,y} \cdot \int_{z_b}^{\eta} g z \tilde{\rho} \mathbf{v} dz = \int_{z_b}^{\eta} g \tilde{\rho} u_3 dz, \tag{19}$$

which is the integrated local conservation of gravitational potential energy.

Regarding constitutive equations, we assume that the fluid is Newtonian i.e. the viscous part of the Cauchy stress depends linearly on the velocity and is given by

$$\sigma = \xi \nabla \cdot \mathbf{U}\, \mathbb{I} + 2\mu D(\mathbf{U}),$$

where μ is the viscosity coefficient, ξ is the second viscosity and $D(\mathbf{U}) = (\nabla\mathbf{U} + (\nabla\mathbf{U})^T)/2$. The heat flux $Q_{\tilde{T}}$ obeys the Fourier law $Q_{\tilde{T}} = -\tilde{\lambda}\nabla\tilde{T}$, which explains the name "Navier-Stokes-Fourier" which is often given to system (15)–(17), $\tilde{\lambda}$ being the heat conductivity.

Among the thermodynamic variables $\tilde{\rho}, \tilde{p}, \tilde{T}, \tilde{e}$, only two of them are independent. Indeed, we have a state law under the form

$$f(\tilde{\rho}, \tilde{T}, \tilde{p}) = 0, \tag{20}$$

where f is a real valued function. We give some examples below. Moreover, the thermodynamic variables are linked by the Gibbs identity

$$d\tilde{e} = \frac{\tilde{p}}{\tilde{\rho}^2} d\tilde{\rho} + \tilde{T} ds, \tag{21}$$

where s is the specific entropy of the fluid. Classically, in order to have a convenient entropy dissipation one has to assume that $-s$ is a convex function of $1/\tilde{\rho}, \tilde{e}$.

2.1.1 Boundary Conditions at the Bottom

Let $\mathbf{n}_b$ and $\mathbf{n}_s$ be the unit outward normals at the bottom and at the free surface respectively, defined by (see Fig. 2)

$$\mathbf{n}_b = \frac{1}{\sqrt{1+|\nabla_{x,y} z_b|^2}} \begin{pmatrix} \nabla_{x,y} z_b \\ -1 \end{pmatrix}, \qquad \mathbf{n}_s = \frac{1}{\sqrt{1+|\nabla_{x,y} \eta|^2}} \begin{pmatrix} -\nabla_{x,y} \eta \\ 1 \end{pmatrix}.$$

On the bottom we prescribe an impermeability condition

$$\mathbf{U} \cdot \mathbf{n}_b = 0, \tag{22}$$

and a friction condition given e.g. by a Navier law

$$\big((\sigma - p\mathbb{I}) \cdot \mathbf{n}_b\big) \cdot \mathbf{t}_i = -\kappa \mathbf{U} \cdot \mathbf{t}_i, \quad i = 1, 2, \tag{23}$$

with κ a Navier coefficient and $(\mathbf{t}_i, i = 1, 2)$ two tangential vectors.

Remark 2 The formulation of the two boundary conditions (22), (23) means that the fluid remains in contact with the topography. Besides, we assume throughout the paper that the total pressure remains non-negative.

2.1.2 Boundary Conditions at Free Surface

On the free surface $z = \eta(t, x, y)$, we use the kinematic boundary condition

$$\frac{\partial \eta}{\partial t} + \mathbf{v}(t, x, y, \eta) \cdot \nabla_{x,y} \eta - u_3(t, x, y, \eta) = 0, \tag{24}$$

and the no stress condition

$$(\sigma - p\mathbb{I}) \cdot \mathbf{n}_s = -p^a(t, x, y)\mathbf{n}_s + W(t, x, y)\mathbf{t}_s, \tag{25}$$

where $p^a(t, x, y)$, $W(t, x, y)$ are two given external forcings, p^a (resp. W) mimics the effects of the atmospheric pressure (resp. the wind blowing at the free surface) and $\mathbf{t}_s$ is a given unit horizontal vector. Throughout the paper we assume $p^a = cst$, $W = 0$.

2.1.3 Boundary Conditions for the Temperature

The heat flux in Eq. (17) requires to define boundary conditions for the temperature. Moreover when the state law (20) will be precised, the definition of the temperature at each boundary will be mandatory. We can choose either Neumann or Dirichlet conditions namely at the bottom

$$\lambda \frac{\partial \tilde{T}}{\partial \mathbf{n}_b} = F\tilde{T}_b^0, \tag{26}$$

or

$$\tilde{T}_b = \tilde{T}_b^0, \tag{27}$$

and at the free surface

$$\lambda \frac{\partial \tilde{T}}{\partial \mathbf{n}_s} = F\tilde{T}_s^0, \tag{28}$$

or

$$\tilde{T}_s = \tilde{T}_s^0, \tag{29}$$

where $F\tilde{T}_b^0$, $F\tilde{T}_s^0$ are two given temperature fluxes and $\tilde{T}_b^0$, T_s^0 are two given temperatures.

2.2 Thermodynamic Considerations

In the following proposition, we propose a formulation of the compressible Euler system—corresponding to the system (15)–(17) with $\lambda = \xi = \mu = 0$—where the acoustic speed explicitly appears. The system is deduced from the compressible Navier-Stokes system (15)–(17) with the boundary conditions (22)–(25).

Proposition 1 *Considering a state law under the form*

$$\tilde{p} = f(\tilde{\rho}, \tilde{T}), \tag{30}$$

the compressible Euler system can be rewritten under the form

$$\frac{\partial \tilde{\rho}}{\partial t} + \nabla \cdot (\tilde{\rho}\mathbf{U}) = 0, \tag{31}$$

$$\frac{\partial (\tilde{\rho}\mathbf{U})}{\partial t} + \nabla \cdot (\tilde{\rho}\mathbf{U} \otimes \mathbf{U}) + \nabla \tilde{p} = \tilde{\rho}\mathbf{g}, \tag{32}$$

$$\frac{\partial (\tilde{\rho}\tilde{p})}{\partial t} + \nabla \cdot (\tilde{\rho}\mathbf{U}\tilde{p}) + \tilde{\rho}^2\tilde{c}^2 \nabla \cdot \mathbf{U} = 0, \tag{33}$$

where the sound speed $\tilde{c}$ is defined below by (47).

The system (31)–(33) is completed with the boundary conditions (22), (24) and (25) which becomes

$$\tilde{p}(t, x, y, \eta) = p^a(t, x, y) = cst. \tag{34}$$

Smooth solutions of the system (31)–(33) *satisfy the energy balance*

$$\frac{\partial}{\partial t}\left(\tilde{\rho}\frac{|\mathbf{U}|^2}{2}+\tilde{\rho}\tilde{e}\right)+\nabla\cdot\left(\left(\tilde{\rho}\frac{|\mathbf{U}|^2}{2}+\tilde{\rho}\tilde{e}+\tilde{p}\right)\mathbf{U}\right)=\tilde{\rho}\mathbf{g}\cdot\mathbf{U}, \tag{35}$$

where the internal energy $\tilde{e}$ *satisfies the equation*

$$\frac{\partial(\tilde{\rho}\tilde{e})}{\partial t}+\nabla\cdot(\tilde{\rho}\tilde{e}\mathbf{U})=-\tilde{p}\nabla\cdot\mathbf{U}. \tag{36}$$

Notice that in this proposition we have kept the same notations even if we have switched from the Navier-Stokes to the Euler system.

For the Euler system (31)–(33)—and also for the Navier-Stokes system—a crucial point is the duality relation between the gradient and divergence operators which writes

$$\int_{\Omega\times[z_b,\eta]}\tilde{p}\nabla.\mathbf{V}d\mathbf{x}dz=\int_{\partial(\Omega\times[z_b,\eta])}\tilde{p}\mathbf{V}.\mathbf{n}ds-\int_{\Omega\times[z_b,\eta]}\mathbf{V}.\nabla\tilde{p}d\mathbf{x}dz. \tag{37}$$

It will be important to have, in the shallow water context, a relation analogous to (37), see (95) below.

Proof ***(Proposition*** *1)* The main point of this proof is the derivation of Eq. (33).

Taking the scalar product of Eq. (16) by $\mathbf{U}$ yields the kinetic energy equation

$$\frac{\partial}{\partial t}\left(\tilde{\rho}\frac{|\mathbf{U}|^2}{2}\right)+\nabla\cdot\left(\left(\tilde{\rho}\frac{|\mathbf{U}|^2}{2}+\tilde{p}-\sigma\right)\mathbf{U}\right)=\tilde{p}\nabla\cdot\mathbf{U}-\sigma:D(\mathbf{U})+\tilde{\rho}\mathbf{g}\cdot\mathbf{U}. \tag{38}$$

Subtracting (38) to (17) gives the equation for the internal energy

$$\frac{\partial(\tilde{\rho}\tilde{e})}{\partial t}+\nabla\cdot(\tilde{\rho}\tilde{e}\mathbf{U})=-\tilde{p}\nabla\cdot\mathbf{U}+\sigma:D(\mathbf{U})-\nabla\cdot Q_{\tilde{T}}, \tag{39}$$

or equivalently

$$\tilde{\rho}\frac{D\tilde{e}}{Dt}=-\tilde{p}\nabla\cdot\mathbf{U}+\sigma:D(\mathbf{U})-\nabla\cdot Q_{\tilde{T}}, \tag{40}$$

with the classical notation $D/Dt\equiv\partial/\partial t+\mathbf{U}\cdot\nabla$. We can write the continuity equation (15) as

$$\tilde{\rho}\frac{D\tilde{\rho}}{Dt}+\tilde{\rho}^2\nabla\cdot\mathbf{U}=0. \tag{41}$$

With the thermodynamic relation (21) one can write $ds=d\tilde{e}/\tilde{T}-(\tilde{p}/\tilde{T}\tilde{\rho}^2)d\tilde{\rho}$, thus multiplying (40) by $1/\tilde{T}$ and (41) by $-\tilde{p}/\tilde{T}\tilde{\rho}^2$ we obtain

$$\tilde{\rho}\frac{Ds}{Dt} = \frac{1}{\tilde{T}}\sigma : D(\mathbf{U}) - \frac{1}{\tilde{T}}\nabla\cdot Q_{\tilde{T}}. \tag{42}$$

This can be written also

$$\frac{\partial(\tilde{\rho}s)}{\partial t} + \nabla\cdot(\tilde{\rho}s\mathbf{U}) = \frac{1}{\tilde{T}}\sigma : D(\mathbf{U}) - \nabla\cdot\frac{Q_{\tilde{T}}}{\tilde{T}} - Q_{\tilde{T}}\cdot\frac{\nabla\tilde{T}}{\tilde{T}^2}, \tag{43}$$

which gives the increase with time of $\int \tilde{\rho}s$, the second principle of thermodynamics.

The state law (20) plays the role of a closure relation. When written under the form (30), it allows to write

$$d\tilde{p} = \left(\frac{\partial\tilde{p}}{\partial\tilde{\rho}}\right)_{\tilde{T}} d\tilde{\rho} + \left(\frac{\partial\tilde{p}}{\partial\tilde{T}}\right)_{\tilde{\rho}} d\tilde{T}. \tag{44}$$

Therefore, from Eqs. (44), (42) and using Eqs. (41) we get

$$\tilde{\rho}\frac{D\tilde{p}}{Dt} + \tilde{\rho}^2\left(\left(\frac{\partial\tilde{p}}{\partial\tilde{\rho}}\right)_{\tilde{T}} + \left(\frac{\partial\tilde{p}}{\partial\tilde{T}}\right)_{\tilde{\rho}}\left(\frac{\partial\tilde{T}}{\partial\tilde{\rho}}\right)_{s}\right)\nabla\cdot\mathbf{U} = \frac{1}{\tilde{T}}\left(\frac{\partial\tilde{p}}{\partial\tilde{T}}\right)_{\tilde{\rho}}\left(\frac{\partial\tilde{T}}{\partial s}\right)_{\tilde{\rho}}\left(\sigma : D(\mathbf{U}) - \nabla\cdot Q_{\tilde{T}}\right) \tag{45}$$

Using the chain rule this can be written

$$\tilde{\rho}\frac{D\tilde{p}}{Dt} + \tilde{\rho}^2\tilde{c}^2\nabla\cdot\mathbf{U} = \frac{1}{\tilde{T}}\left(\frac{\partial\tilde{p}}{\partial s}\right)_{\tilde{\rho}}\left(\sigma : D(\mathbf{U}) - \nabla\cdot Q_{\tilde{T}}\right), \tag{46}$$

with the sound speed $\tilde{c}$ given by

$$\tilde{c}^2 = \left(\frac{\partial\tilde{p}}{\partial\tilde{\rho}}\right)_{\tilde{T}} + \left(\frac{\partial\tilde{p}}{\partial\tilde{T}}\right)_{\tilde{\rho}}\left(\frac{\partial\tilde{T}}{\partial\tilde{\rho}}\right)_{s} = \left(\frac{\partial\tilde{p}}{\partial\tilde{\rho}}\right)_{s}. \tag{47}$$

And coupled with (15), Eq. (46) writes

$$\frac{\partial(\tilde{\rho}\tilde{p})}{\partial t} + \nabla\cdot(\tilde{\rho}\mathbf{U}\tilde{p}) + \tilde{\rho}^2\tilde{c}^2\nabla\cdot\mathbf{U} = \frac{1}{\tilde{T}}\left(\frac{\partial\tilde{p}}{\partial s}\right)_{\tilde{\rho}}\left(\sigma : D(\mathbf{U}) - \nabla\cdot Q_{\tilde{T}}\right), \tag{48}$$

Equations (15), (16) and (48) with $\lambda = \xi = \mu = 0$ give Eqs. (31)–(33).

Using (36), taking the scalar product of (32) with $\mathbf{U}$ and after simple computations, we obtain the energy balance (35). □

Remark 3 Whereas, Eq. (31) expresses the local mass conservation, the volume variations can be related to the temperature variations. Indeed, since

$$d\tilde{T} = \left(\frac{\partial \tilde{T}}{\partial \tilde{\rho}}\right)_s d\tilde{\rho} + \left(\frac{\partial \tilde{T}}{\partial s}\right)_{\tilde{\rho}} ds,$$

using relations (42) and (41) we get the following equation governing the temperature

$$\tilde{\rho}\frac{D\tilde{T}}{Dt} + \tilde{\rho}^2\left(\frac{\partial \tilde{T}}{\partial \tilde{\rho}}\right)_s \nabla \cdot \mathbf{U} = \frac{1}{\tilde{T}}\left(\frac{\partial \tilde{T}}{\partial s}\right)_{\tilde{\rho}} \left(\sigma : D(\mathbf{U}) - \nabla \cdot Q_{\tilde{T}}\right). \tag{49}$$

2.3 Acoustic Waves and Water Waves

The system (31)–(33)—completed with the boundary conditions (22), (24) and (34)—is a compressible model with a free surface and hence acoustic and water waves can propagate.

Let us define $\hat{p}$ by

$$\tilde{p} = p^a + \int_z^\eta \tilde{\rho} g dz + \hat{p},$$

with $p^a = cst$ thus $\hat{p}$ denotes the non-gravitational part of the pressure. Then the system (31)–(33) with (22) and (24) also writes

$$\frac{\partial \tilde{\rho}}{\partial t} + \nabla \cdot (\tilde{\rho}\mathbf{U}) = 0, \tag{50}$$

$$\frac{\partial \mathbf{U}}{\partial t} + (\mathbf{U} \cdot \nabla)\mathbf{U} + \frac{1}{\tilde{\rho}}\nabla \hat{p} + \frac{1}{\tilde{\rho}}\nabla \int_z^\eta \tilde{\rho} g dz_1 = \mathbf{g}, \tag{51}$$

$$\frac{\partial}{\partial t}\left(\int_z^\eta \tilde{\rho} g dz + \hat{p}\right) + \mathbf{U} \cdot \nabla \left(\int_z^\eta \tilde{\rho} g dz + \hat{p}\right) + \tilde{\rho}\tilde{c}^2 \nabla \cdot \mathbf{U} = 0, \tag{52}$$

$$\frac{\partial h}{\partial t} + \mathbf{v}_s \cdot \nabla_{x,y} h = u_{3,s}, \tag{53}$$

where the subscript corresponds to the value of the free surface $z = \eta$.

Assuming a flat bottom and in a two dimensional setting (x, z), the system (50)–(53) has the following compact formulation

$$M\frac{\partial Y}{\partial t} + A_x\frac{\partial Y}{\partial x} + A_z\frac{\partial Y}{\partial z} = S, \tag{54}$$

with

$$Y = \begin{pmatrix} \tilde{\rho} \\ u_1 \\ u_3 \\ \hat{p} \\ h \end{pmatrix}, \quad A_x = \begin{pmatrix} u_1 & \tilde{\rho} & 0 & 0 & 0 \\ \frac{g(h-z)}{\tilde{\rho}} & u_1 & 0 & \frac{1}{\tilde{\rho}} & g \\ 0 & 0 & u_1 & 0 & 0 \\ g(h-z)u_1 & \tilde{\rho}\tilde{c}^2 & 0 & u_1 & g\tilde{\rho}u_1 \\ 0 & 0 & 0 & 0 & u_{1,s} \end{pmatrix},$$

$$A_z = \begin{pmatrix} u_3 & 0 & \tilde{\rho} & 0 & 0 \\ 0 & u_3 & 0 & 0 & 0 \\ 0 & 0 & u_3 & \frac{1}{\tilde{\rho}} & 0 \\ 0 & 0 & \tilde{\rho}\tilde{c}^2 & u_3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ g(h-z) & 0 & 0 & 1 & \tilde{\rho}g \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and

$$S = \begin{pmatrix} 0 \\ \frac{g}{\tilde{\rho}} \frac{\partial}{\partial x} \int_z^h \big(\tilde{\rho}(t,x,z) - \tilde{\rho}(t,x,z_1)\big) dz_1 \\ 0 \\ g \frac{\partial}{\partial t} \int_z^h \big(\tilde{\rho}(t,x,z) - \tilde{\rho}(t,x,z_1)\big) dz_1 - g u_1 \frac{\partial}{\partial x} \int_z^h \big(\tilde{\rho}(t,x,z) - \tilde{\rho}(t,x,z_1)\big) dz_1 + g\tilde{\rho}u_3 \\ u_{3,s} \end{pmatrix}.$$

Considering we are in a shallow water context, we can further assume

$$\frac{\partial \tilde{\rho}}{\partial z} = \frac{\partial u_1}{\partial z} = 0, \tag{55}$$

then the system (54) reduces to

$$M^{sw} \frac{\partial Y}{\partial t} + A_x^{sw} \frac{\partial Y}{\partial x} + A_z \frac{\partial Y}{\partial z} = B, \tag{56}$$

with

$$A_x^{sw} = \begin{pmatrix} u_1 & \tilde{\rho} & 0 & 0 & 0 \\ \frac{g(h-z)}{\tilde{\rho}} & u_1 & 0 & \frac{1}{\tilde{\rho}} & g \\ 0 & 0 & u_1 & 0 & 0 \\ ghu_1 & \tilde{\rho}\big(\tilde{c}^2 + gz\big) & 0 & u_1 & g\tilde{\rho}u_1 \\ 0 & 0 & 0 & 0 & u_1 \end{pmatrix}, \quad M^{sw} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ gh & 0 & 0 & 1 & \tilde{\rho}g \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The eigenvalues of the matrix $aA_x^{sw} + bA_z^{sw}$ for $(a,b) \in \mathbb{R}^2$ cannot be easily computed explicitely but the following result holds.

Proposition 2 *The eigenvalues of the matrix* $(M^{sw})^{-1}(aA_x^{sw} + bA_z^{sw})$ *with* $u_1 = u_3 = 0$, *are given by*

$$0, \pm\frac{1}{2}\sqrt{2C_1 \pm 2\sqrt{C_2 + C_3}}, \tag{57}$$

with

$$C_1 = \tilde{c}^2(a^2+b^2) - b^2gh, \quad C_2 = (a^2+b^2)^2\tilde{c}^4 + b^2g^2(8a^2hz - 4a^2z^2 + b^2h^2),$$
$$C_3 = -2b^2(3a^2+b^2)\tilde{c}^2gh.$$

Proof ***(Proposition 2)*** The proof relies on simple computations that are not presented here. □

Remark 4 Notice that the quantities C_2 is non-negative whereas C_3 is non-positive. In the situation where $\tilde{c}^2 \geq gh$ that is encountered in practice, then $C_1 \geq 0$, $C_2 + C_3 \geq 0$ and $C_1 - \sqrt{C_2 + C_3} \geq 0$ therefore the system has 4 real eigenvalues. But when $\tilde{c}^2 \leq gh$—corrresponding to a less realistic situation—then complex eigenvalues could appear.

Notice also that we are considering situations where $\tilde{c} \gg 1$ (see Eq. (68) below), hence the eigenvalues defined by (57) satisfy the estimates

$$0, \quad \pm\tilde{c}\sqrt{a^2+b^2} + O\Big(\frac{1}{\tilde{c}}\Big), \quad \pm\sqrt{gh}\,\frac{ab}{\sqrt{a^2+b^2}} + O\Big(\frac{1}{\tilde{c}^2}\Big), \tag{58}$$

where the second ones correspond to acoustic waves and the third one to surface waves. The first eigenvalue is zero because of the linearization of the velocity field.

2.4 Sound Speed for Sea Water

In this paragraph, we are going to precise the expression of the sound speed $\tilde{c}$ defined by (47) in the particular case of sea water.

We start from an expression of the state law (30) given by [29] under the form

$$\tilde{\rho} = \tilde{\rho}(\tilde{p}, \tilde{T}) = \frac{\tilde{\rho}_0(\tilde{T})}{1 - \frac{\varepsilon}{\rho_0}\tilde{p}} = \frac{\rho_0\tilde{\rho}_0(\tilde{T})}{\rho_0 - \varepsilon\tilde{p}}, \tag{59}$$

ρ_0 and ε being two constants with $\varepsilon \ll 1$ and where

$$\tilde{\rho}_0(\tilde{T}) = \rho_0 + a(\tilde{T} - T_0)^2, \tag{60}$$

with $T_0 = 4$ °C, $\rho_0 = 9999.7$ kg.m^{-3}, $a = -6.63\ 10^{-3}$ kg.m^{-3}.K^{-2}. Notice that from the state law (59) we have

$$\varepsilon\tilde{p} = o(1). \tag{61}$$

When multiplied by ε, the relation (21) is compatible with the scaling (61) if

$$\varepsilon\tilde{e} = o(1), \quad \text{and} \quad \varepsilon s = o(1). \tag{62}$$

Rewritting (21) under the form

$$d(\tilde{e} - \tilde{T}s) = -sd\tilde{T} + \frac{\tilde{p}}{\tilde{\rho}^2}d\tilde{\rho}, \tag{63}$$

the Schwarz theorem applied to (63) gives the equality

$$-\left(\frac{\partial s}{\partial \tilde{\rho}}\right)_{\tilde{T}} = \left(\frac{\partial(\tilde{p}/\tilde{\rho}^2)}{\partial \tilde{T}}\right)_{\tilde{\rho}},$$

and using the expression of $\tilde{p}$ given by (59) we obtain

$$\left(\frac{\partial s}{\partial \tilde{\rho}}\right)_{\tilde{T}} = \frac{\rho_0 \tilde{\rho}_0'(\tilde{T})}{\varepsilon \tilde{\rho}^3}.$$

An integration of the previous relation gives

$$s = s_0(\tilde{T}) - \frac{\rho_0 \tilde{\rho}_0'(\tilde{T})}{2\varepsilon \tilde{\rho}^2}, \tag{64}$$

and from (64) we obtain

$$ds = \left(s_0'(\tilde{T}) - \frac{\rho_0 \tilde{\rho}_0''(\tilde{T})}{2\varepsilon \tilde{\rho}^2}\right) d\tilde{T} + \frac{\rho_0 \tilde{\rho}_0'(\tilde{T})}{\varepsilon \tilde{\rho}^3} d\tilde{\rho},$$

leading to

$$\left(\frac{\partial \tilde{T}}{\partial \tilde{\rho}}\right)_s = -\frac{2}{\tilde{\rho}} \frac{\rho_0 \tilde{\rho}_0'(\tilde{T})}{2\varepsilon \tilde{\rho}^2 s_0'(\tilde{T}) - \rho_0 \tilde{\rho}_0''(\tilde{T})}. \tag{65}$$

Using (59), (65) we obtain the expression for the sound speed $\tilde{c}$ defined by (47) under the form

$$\tilde{c}^2 = \frac{\rho_0 \tilde{\rho}_0(\tilde{T})}{\varepsilon \tilde{\rho}^2} + \frac{\rho_0 \tilde{\rho}_0'(\tilde{T})}{\varepsilon \tilde{\rho}} \frac{2}{\tilde{\rho}} \frac{\rho_0 \tilde{\rho}_0'(\tilde{T})}{2\varepsilon \tilde{\rho}^2 s_0'(\tilde{T}) - \rho_0 \tilde{\rho}_0''(\tilde{T})}. \tag{66}$$

Because of the estimate (62) concerning the entropy, a possible choice is $s_0'(\tilde{T}) = 0$ leading to the expression

$$\tilde{\rho}^2 \tilde{c}^2 = \frac{\rho_0}{\varepsilon} \left(\tilde{\rho}_0(\tilde{T}) - 2\frac{\left(\tilde{\rho}_0'(\tilde{T})\right)^2}{\tilde{\rho}_0''(\tilde{T})}\right). \tag{67}$$

Moreover, since $\varepsilon \ll 1$ and $a \ll 1$, we have $\tilde{\rho}(\tilde{T}) \approx \rho_0$ leading to an expression for the sound speed under the form

$$\tilde{\rho}^2\tilde{c}^2 \approx \frac{\rho_0^2}{\varepsilon}\left(1 - 4a\frac{\left(\tilde{T} - T_0\right)^2}{\rho_0}\right) \approx \frac{\rho_0^2}{\varepsilon} = \rho_0^2 c^2. \tag{68}$$

Using the assumption (68), it is possible to combine Eq. (33) with Eq. (31) to obtain

$$\frac{\partial}{\partial t}\left(\tilde{\rho}\frac{\tilde{p}^2}{2\rho_0^2c^2}\right) + \nabla\cdot\left(\tilde{\rho}\frac{\tilde{p}^2}{2\rho_0^2c^2}\mathbf{U}\right) + \tilde{p}\nabla\cdot\mathbf{U} = 0, \tag{69}$$

giving a formulation of the energy balance (35) under the form

$$\frac{\partial}{\partial t}\left(\tilde{\rho}\frac{|\mathbf{U}|^2}{2} + \tilde{\rho}\frac{\tilde{p}^2}{2\rho_0^2c^2}\right) + \nabla\cdot\left(\left(\tilde{\rho}\frac{|\mathbf{U}|^2}{2} + \tilde{\rho}\frac{\tilde{p}^2}{2\rho_0^2c^2} + \tilde{p}\right)\mathbf{U}\right) = \tilde{\rho}\mathbf{g}\cdot\mathbf{U}. \tag{70}$$

Hence, when $\tilde{\rho}^2\tilde{c}^2 = \rho_0^2c^2 = cst$, the internal energy corresponds to $\frac{\tilde{p}^2}{2\rho_0^2c^2}$.

2.5 A Shallow Water Approximation of the Compressible Euler System

For free surface flows, the vertical direction plays a particular role since it corresponds to the direction of the gravity. Moreover the fluid domain, in our case, is thin in this direction and it is natural to perform a depth averaging of system (31)–(33) together with some approximations. Then, the two following propositions hold.

Remark 5 Notice that whereas the models described in the sequel differ from the model (9)–(11), for the sake of simplicity, we have kept the same notations for some of the variables of each model.

Proposition 3 *Assuming $\tilde{\rho}\tilde{c}$ is constant, see (68), a shallow water approximation of the compressible Euler system (31)–(33) completed with (22), (24) leads to the model*

$$\frac{\partial\eta}{\partial t} + \mathbf{u}\cdot\nabla_0(\eta + z_b) = \gamma w, \tag{71}$$

$$\frac{\partial(\rho h)}{\partial t} + \nabla_0\cdot(\rho h\mathbf{u}) = 0, \tag{72}$$

$$\frac{\partial(\rho hu)}{\partial t} + \frac{\partial}{\partial x}\left(\rho hu^2 + \frac{\rho g}{2}h^2 + hP\right) + \frac{\partial(\rho huv)}{\partial y} = -(\rho gh + \frac{\gamma^2}{2}P)\frac{\partial z_b}{\partial x}, \tag{73}$$

$$\frac{\partial(\rho hv)}{\partial t} + \frac{\partial(\rho huv)}{\partial x} + \frac{\partial}{\partial y}\left(\rho hv^2 + \frac{\rho g}{2}h^2 + hP\right) = -(\rho gh + \frac{\gamma^2}{2}P)\frac{\partial z_b}{\partial y}, \tag{74}$$

$$\frac{\partial(\rho hw)}{\partial t} + \nabla_0\cdot(\rho hw\mathbf{u}) = \gamma P, \tag{75}$$

$$\frac{\partial}{\partial t}\left(\rho hP + \frac{\rho^2gh^2}{2}\right) + \nabla_0\cdot\left(\left(\rho hP + \frac{\rho^2gh^2}{2}\right)\mathbf{u}\right) + \rho_0^2c^2\mathrm{div}_{sw}^{\gamma}(\mathbf{u}) = 0, \tag{76}$$

where ρ,$\mathbf{u} = (u, v, w)^T$, P represent respectively a density, a velocity vector, and a pressure term, all functions of (t, x, y), and γ a parameter.

The model described in Proposition 3 satisfies an energy balance described in the following proposition.

Proposition 4 *Smooth solutions of the system (71)–(76) satisfy the energy balance*

$$\frac{\partial}{\partial t}\Big(\frac{\rho h}{2}|\mathbf{u}|^2 + \frac{1}{2}\left(2 - \frac{\gamma^2}{2}\right)\rho g h z_b + \rho h e\Big) + \nabla_0 \cdot \Big(\mathbf{u}(\frac{\rho h}{2}|\mathbf{u}|^2 + \frac{1}{2}\left(2 - \frac{\gamma^2}{2}\right)\rho g h z_b + \rho h e + \frac{\rho g}{2}h^2 + hP)\Big) = -\frac{\gamma}{2}\rho g h w, \quad (77)$$

where e is defined by

$$e = \frac{1}{2\rho_0^2 c^2}\Big(P + \frac{\rho g h}{2}\Big)^2. \quad (78)$$

And Eq. (77) can be written in a conservative form

$$\frac{\partial}{\partial t}\Big(\frac{\rho h}{2}|\mathbf{u}|^2 + \rho g\frac{h(\eta + z_b)}{2} + \frac{1}{2}\left(2 - \frac{\gamma^2}{2}\right)\rho g h z_b + \rho h e\Big) + \nabla_0 \cdot \Big(\mathbf{u}(\frac{\rho h}{2}|\mathbf{u}|^2 + \rho g\frac{h(\eta + z_b)}{2} + \frac{1}{2}\left(2 - \frac{\gamma^2}{2}\right)\rho g h z_b + \rho h e + \frac{\rho g}{2}h^2 + hP)\Big) = 0. \quad (79)$$

In other words, (78) gives a shallow water version of the internal energy defined after (36).

The two operators $\mathrm{div}_{sw}^{\gamma}$ and ∇_{sw}^{γ} defined by (12), (13) appear in Eqs. (71)–(76) so that we can rewrite this system in a more compact form

$$\frac{\partial \eta}{\partial t} + \mathbf{u} \cdot \nabla_0(\eta + z_b) = \gamma w, \quad (80)$$

$$\frac{\partial(\rho h)}{\partial t} + \nabla_0 \cdot (\rho h \mathbf{u}) = 0, \quad (81)$$

$$\frac{\partial(\rho h \mathbf{u})}{\partial t} + \nabla_0 \cdot (\rho h \mathbf{u} \otimes \mathbf{u}) + \nabla_0(\frac{\rho g}{2}h^2) + \nabla_{sw}^{\gamma} P = -gh\nabla_0 z_b, \quad (82)$$

$$\frac{\partial}{\partial t}\big(\rho h P + \frac{\rho^2 g h^2}{2}\big) + \nabla_0 \cdot \Big(\big(\rho h P + \frac{\rho^2 g h^2}{2}\big)\mathbf{u}\Big) + \rho_0^2 c^2 \mathrm{div}_{sw}^{\gamma}(\mathbf{u}) = 0, \quad (83)$$

An important point is that whatever the value of γ, these operators satisfy the duality relation

$$\int_\Omega \nabla_{sw}^{\gamma}(f) \cdot \mathbf{w} d\mathbf{x} = -\int_\Omega \mathrm{div}_{sw}^{\gamma}(\mathbf{w}) f d\mathbf{x} + \int_\Gamma h f \mathbf{w} \cdot \mathbf{n} ds, \quad (84)$$

where the vector $\mathbf{n} = (n_x, n_y, 0)^T$ is the outward unit normal vector to the boundary Γ, see Fig. 1. In Eq. (84), f and $\mathbf{w}$ belong to suitable function spaces that will be precised later. Notice that, we also have a local form of (84) under the form

$$\nabla_{sw}^{\gamma}(f) \cdot \mathbf{w} = \nabla_0 \cdot (hf\mathbf{w}) - \mathrm{div}_{sw}^{\gamma}(\mathbf{w})\, f. \tag{85}$$

Proof ***(Proposition 3)*** It is easy to see (cf. [21, Lemma 2.1]) that a depth averaging of the compressible Euler system with gravity and free surface (31)–(33)—with, according to Sect. 2.4, $\tilde{\rho}\tilde{c} \approx \rho_0 c$—completed with the boundary conditions (22), (24), (34) leads to

$$\frac{\partial}{\partial t}\int_{z_b}^{\eta} \tilde{\rho} dz + \nabla_{x,y} \cdot \int_{z_b}^{\eta} \tilde{\rho}\mathbf{v} dz = 0, \tag{86}$$

$$\frac{\partial}{\partial t}\int_{z_b}^{\eta} \tilde{\rho}\mathbf{v} dz + \nabla_{x,y} \cdot \int_{z_b}^{\eta} \tilde{\rho}\mathbf{v} \otimes \mathbf{v} dz + \nabla_{x,y} \int_{z_b}^{\eta} \tilde{p} dz = \tilde{p}(t, \mathbf{x}, z_b(\mathbf{x}))\nabla_{x,y} z_b, \tag{87}$$

$$\frac{\partial}{\partial t}\int_{z_b}^{\eta} \tilde{\rho} u_3 dz + \nabla_{x,y} \cdot \int_{z_b}^{\eta} \tilde{\rho} u_3 \mathbf{v} dz = \tilde{p}(t, \mathbf{x}, z_b(\mathbf{x})) - \int_{z_b}^{\eta} \tilde{\rho} g dz, \tag{88}$$

$$\frac{\partial}{\partial t}\int_{z_b}^{\eta} \tilde{\rho}\tilde{p} dz + \nabla_{x,y} \cdot \int_{z_b}^{\eta} \tilde{\rho}\tilde{p}\mathbf{v} dz + \rho_0^2 c^2 \int_{z_b}^{\eta} \nabla \cdot \mathbf{U} dz = 0. \tag{89}$$

As in [2] we are now going to make some assumptions concerning the variations along the vertical axis of the velocity field $\mathbf{U}$, the density $\tilde{\rho}$ and of the pressure $\tilde{p}$. In order to be consistent with the shallow water assumption, the choice

$$u_1(t, \mathbf{x}, z) = u(t, \mathbf{x}), \quad u_2(t, \mathbf{x}, z) = v(t, \mathbf{x}), \quad \tilde{\rho}(t, \mathbf{x}, z) = \rho(t, \mathbf{x}), \tag{90}$$

is natural since it consists in assimilating the horizontal velocity field and the density with their vertical means. For the velocity u_3 and the pressure $\tilde{p}$, we choose

$$u_3(t, \mathbf{x}, z) = \varphi_\delta\left(\frac{\eta - z}{h}\right) w(t, \mathbf{x}), \tag{91}$$

$$\tilde{p}(t, \mathbf{x}, z) = \rho g(\eta - z) + \psi_\delta\left(\frac{\eta - z}{h}\right) P(t, \mathbf{x}), \tag{92}$$

and the two families of functions $\psi_\delta = \psi_\delta(z)$ and $\varphi_\delta = \varphi_\delta(z)$ satisfy

$$\begin{cases} \int_0^1 \varphi_\delta(z) dz = \int_0^1 \psi_\delta(z) dz = \frac{1}{2}\int_0^1 \varphi_\delta(z)\psi_\delta'(z) dz = 1, \\ \psi_\delta(1) = \delta, \ \psi_\delta(0) = 0, \ \varphi_\delta(1) = 1. \end{cases} \tag{93}$$

Notice that these choices are similar to those in [2]. Figure 2 in [2, paragraph 2.3.2] illustrates the shape of the functions ψ_δ and φ_δ for two typical values of δ namely $\delta = 2$ and $\delta = 3/2$ (corresponding to $\gamma = 2$ and $\gamma = \sqrt{3}$). It appears that the functions ψ_δ and φ_δ do not significantly differ when $\delta = 2$ or when $\delta = 3/2$, the choice $\delta = 2$ corresponding to a linear profile.

The duality relation (37) is a guideline for the definition of the shallow water version of the divergence operator. Therefore, from (90)–(93) and using an integration by parts, we obtain

$$\begin{aligned}\int_{z_b}^{\eta} \psi_\delta\left(\frac{\eta-z}{h}\right)\nabla\cdot\mathbf{U}dz &= \int_{z_b}^{\eta} \psi_\delta\left(\frac{\eta-z}{h}\right)\nabla_0\cdot\mathbf{u}dz + \left[u_3\psi_\delta\left(\frac{\eta-z}{h}\right)\right]_{z_b}^{\eta}\\ &\quad - \int_{z_b}^{\eta} w\varphi_\delta\left(\frac{\eta-z}{h}\right)\frac{\partial}{\partial z}\psi_\delta\left(\frac{\eta-z}{h}\right)dz\\ &= 2w + h\nabla_0\cdot\mathbf{u} - \delta\mathbf{u}\cdot\nabla_0 z_b, \end{aligned} \tag{94}$$

where (22) has been used and allowing us to write

$$\begin{aligned}\int_{z_b}^{\eta}\tilde{p}\nabla\cdot\mathbf{U}dz &\approx \int_{z_b}^{\eta}\left(\frac{1}{h}\int_{z_b}^{\eta} g(\eta-z_1)dz_1 + P\right)\psi_\delta\left(\frac{\eta-z}{h}\right)\nabla\cdot\mathbf{U}dz\\ &= \left(\frac{\rho g h}{2}+P\right)\int_{z_b}^{\eta}\psi_\delta\left(\frac{\eta-z}{h}\right)\nabla\cdot\mathbf{U}dz\\ &= \left(\frac{\rho g h}{2}+P\right)\left(2w + h\nabla_0\cdot\mathbf{u} - \delta\mathbf{u}\cdot\nabla_0 z_b\right). \end{aligned} \tag{95}$$

The computations (94), (95) are used to approximate the last term in Eq. (89) under the form

$$\int_{z_b}^{\eta}\nabla\cdot\mathbf{U}dz \approx 2w + h\nabla_0\cdot\mathbf{u} - \delta\mathbf{u}\cdot\nabla_0 z_b.$$

And with the choices (90)–(93), the system (86)–(89) writes

$$\frac{\partial(\rho h)}{\partial t} + \nabla_0\cdot(\rho h\mathbf{u}) = 0, \tag{96}$$

$$\frac{\partial(\rho h u)}{\partial t} + \frac{\partial}{\partial x}\left(\rho h u^2 + \frac{\rho g}{2}h^2 + hP\right) + \frac{\partial(\rho h u v)}{\partial y} = -(\rho g h + \delta P)\frac{\partial z_b}{\partial x}, \tag{97}$$

$$\frac{\partial(\rho h v)}{\partial t} + \frac{\partial(\rho h u v)}{\partial x} + \frac{\partial}{\partial y}\left(\rho h v^2 + \frac{\rho g}{2}h^2 + hP\right) = -(\rho g h + \delta P)\frac{\partial z_b}{\partial y}, \tag{98}$$

$$\frac{\partial(\rho h w)}{\partial t} + \nabla_0\cdot(\rho h w\mathbf{u}) = \delta P, \tag{99}$$

$$\begin{aligned}&\frac{\partial}{\partial t}\left(\rho h P + \frac{\rho^2 g h^2}{2}\right) + \nabla_0\cdot\left(\left(\rho h P + \frac{\rho^2 g h^2}{2}\right)\mathbf{u}\right)\\ &\qquad + \rho_0^2 c^2\left(2w + h\nabla_0\cdot\mathbf{u} - \delta\mathbf{u}\cdot\nabla_0 z_b\right) = 0. \end{aligned} \tag{100}$$

Using the choices (90)–(93), Eq. (19) gives

$$\frac{\partial}{\partial t}\left(\frac{\rho h(\eta+z_b)}{2}\right) + \nabla_0\cdot\left(\frac{\rho h(\eta+z_b)}{2}\mathbf{u}\right) = \rho h w, \tag{101}$$

and combining (101) with (96) gives

$$\frac{\partial \eta}{\partial t} + \mathbf{u} \cdot \nabla_0(\eta + z_b) = 2w. \tag{102}$$

Finally, a simple change of variables, namely $w = \gamma \hat{w}/2$ with $\gamma^2 = 2\delta$ in the system (96)–(100), (102) leads to Eqs. (71)–(76) where the symbol $\hat{}$ has been dropped. □

Proof ***(Proposition 4)*** After simple computations and using Eq. (72), Eq. (76) multiplied by $(P + \frac{\rho g h}{2})/(\rho_0^2 c^2)$ gives

$$\frac{\partial}{\partial t}\left(\frac{\rho h}{2\rho_0^2 c^2}\left(P + \frac{\rho g h}{2}\right)^2\right) + \nabla_0 \cdot \left(\frac{\rho h}{2\rho_0^2 c^2}\left(P + \frac{\rho g h}{2}\right)^2 \mathbf{u}\right) + \left(P + \frac{\rho g h}{2}\right) \mathrm{div}_{sw}^{\gamma}(\mathbf{u}) = 0, \tag{103}$$

Now, taking the scalar product of Eqs. (73)–(75) with $\mathbf{u}$, using the duality relation (85) and adding the obtained relation with (103) gives (77).

And the sum of Eq. (77) with (101) multiplied by g—in which the change of variable $w = \gamma \hat{w}/2$ is done and the symbol $\hat{}$ has been dropped—gives (79).

□

2.6 When the Density is Almost Constant

On the one hand, the propagation of acoustic waves requires a compressible medium, on the other hand the variations of the fluid density are often neglected e.g. when considering a linearized version of the Euler system (31)–(33).

In this paragraph, we assume that the variations of the fluid density have little influence over the hydrodynamic regime and the waves propagation, that is not a strong assumption for water, see (60). Nevertheless, it is not possible to simply consider that the density is constant in the considered models. Indeed, the assumption $\rho = cst$ in the 3d case—Eq. (15)—or in the shallow water context—Eq. (101)—leads to a divergence free condition that is not compatible with the equations governing the pressure variations namely Eq. (33) or Eq. (76) in the shallow water regime.

Hence when the variations of the fluid density can be neglected, the Proposition 3 can be reformulated as follows.

Proposition 5 *Assuming the setting of Proposition 3 and neglecting the variation of the fluid density, a shallow water approximation of the compressible Euler system* (31)–(33) *is given by*

$$\frac{\partial h}{\partial t} + \nabla_0 \cdot (h\mathbf{u}) = 0, \tag{104}$$

$$\frac{\partial (h\mathbf{u})}{\partial t} + \nabla_0 \cdot (h\mathbf{u} \otimes \mathbf{u}) + \nabla_0 (\frac{g}{2} h^2) + \nabla_{sw}^{\gamma} p = -gh\nabla_0 z_b, \tag{105}$$

$$\frac{\partial}{\partial t}\big(hp + \frac{gh^2}{2}\big) + \nabla_0 \cdot \Big(\big(hp + \frac{gh^2}{2}\big)\mathbf{u}\Big) + c^2 \mathrm{div}_{sw}^{\gamma}(\mathbf{u}) = 0, \tag{106}$$

where the operators ∇_{sw}^{γ} *and* $\mathrm{div}_{sw}^{\gamma}$ *are defined by* (12), (13).

Proof ***(Proposition 5)*** The model (81)–(83) is nothing else than a rewritting of Eqs. (72)–(76) where the variations of the fluid density are neglected i.e. $\rho \equiv \rho_0$ and we have introduced $p = P/\rho_0$. □

For the model obtained in Proposition 3, Eq. (101) is crucial to obtain an energy balance. In order to obtain an energy balance for the model given in Proposition 5, we introduce a function $\tilde{\zeta} = \tilde{\zeta}(t, \mathbf{x})$ solution of the transport equation

$$\frac{\partial \tilde{\zeta}}{\partial t} + \mathbf{u} \cdot \nabla_0 \tilde{\zeta} = \gamma w, \tag{107}$$

or equivalently using (104)

$$\frac{\partial (h\tilde{\zeta})}{\partial t} + \nabla_0 \cdot (h\tilde{\zeta}\mathbf{u}) = h\gamma w. \tag{108}$$

From the definition (13) and using (104), we can write

$$h\mathrm{div}_{sw}^{\gamma}(\mathbf{u}) = h\gamma w + h\nabla_0 \cdot (h\mathbf{u}) - h\mathbf{u} \cdot \nabla_0 \zeta = h\gamma w - \frac{\partial (h\zeta)}{\partial t} - \nabla_0 \cdot \big(h\zeta \mathbf{u}\big),$$

and hence, Eq. (108) can be written under the form

$$\frac{\partial (h\hat{\zeta})}{\partial t} + \nabla_0 \cdot \big(h\hat{\zeta}\mathbf{u}\big) = h\mathrm{div}_{sw}^{\gamma}(\mathbf{u}), \tag{109}$$

with $\hat{\zeta} = \tilde{\zeta} - h - \gamma^2 z_b/2 = \tilde{\zeta} - \zeta$. Notice that $\tilde{\zeta}$ is an approximation in the constant density case of the variable $\eta + z_b$ governed by Eq. (80). And from Eq. (106), $\hat{\zeta} = O(1/c^2)$. The existence of solution for Eq. (107) has been widely studied, let us mention the contributions of Di Perna and Lions by the means of renormalized solutions [17] and two extensions [16, 26], see also [11]. We assume here that the variables h, $\mathbf{u}$ and the quantity z_b are regular enough so that these existence results are valid.

As already mentionned, the assumption $\rho = cst$ implies that mass and volume are conserved that could be seen as contradictory with the capability of acoustic waves to propagate—with finite speed—since it requires a compressibility in the considered media. But the quantity $\hat{\zeta}$ can also be related to the temperature effects and allows to circumvent this difficulty. More precisely, when $\lambda = \mu = 0$ and assuming

$$\left(\frac{\partial \tilde{T}}{\partial \tilde{\rho}}\right)_s = \tau = cst,$$

Equation (109) looks like a shallow water version of Eq. (49), the variable $\hat{\zeta}$ corresponding to a shallow water approximation of the quantity $\frac{\tilde{T}}{\tau\rho_0}$. Hence the variations of $\hat{\zeta}$ correspond to volume variations and can be assimilated to dilatation effects generated by temperature variations.

Proposition 6 *Smooth solutions of the system* (104)–(106) *satisfy the energy balance*

$$\begin{aligned}\frac{\partial}{\partial t}\Big(\frac{h}{2}|\mathbf{u}|^2 + \frac{g}{2}\left(2-\frac{\gamma^2}{2}\right)hz_b + \frac{h}{2c^2}\big(p+\frac{gh}{2}\big)^2\Big) + \nabla_0\cdot\Big(\mathbf{u}\big(\frac{h}{2}|\mathbf{u}|^2\\ + \frac{g}{2}\left(2-\frac{\gamma^2}{2}\right)hz_b + \frac{h}{2c^2}\big(p+\frac{gh}{2}\big)^2 + \frac{g}{2}h^2 + hp\big)\Big) = -\frac{\gamma h}{2}gw,\end{aligned} \tag{110}$$

that is a shallow water version of Eq. (35)*. Equation* (110) *can be rewritten under a conservative form given by*

$$\begin{aligned}\frac{\partial}{\partial t}\Big(\frac{h}{2}|\mathbf{u}|^2 + g\frac{h(\eta+z_b)}{2} + g\frac{h\hat{\zeta}}{2} + \frac{h}{2c^2}\big(p+\frac{gh}{2}\big)^2\Big) + \nabla_0\cdot\Big(\mathbf{u}\big(\frac{h}{2}|\mathbf{u}|^2\\ + g\frac{h(\eta+z_b)}{2} + \frac{h}{2c^2}\big(p+\frac{gh}{2}\big)^2 + \frac{g}{2}h^2 + g\frac{h\hat{\zeta}}{2} + hp\big)\Big) = 0.\end{aligned} \tag{111}$$

Multiplying Eq. (106) by $p + gh/2$ and after simple computations, we obtain the relation

$$\frac{\partial}{\partial t}\left(\frac{h}{2c^2}\big(p+\frac{gh}{2}\big)^2\right) + \nabla_0\cdot\left(\frac{h}{2c^2}\Big(\big(p+\frac{gh}{2}\big)^2\mathbf{u}\Big)\right) + (p+\frac{gh}{2})\mathrm{div}^{\gamma}_{sw}(\mathbf{u}) = 0. \tag{112}$$

And comparing (112) with (36) we obtain that when the density is kept constant, the internal energy is given by $(p + gh/2)^2/(2c^2)$, see also (78).

Proof ***(Proposition 6)*** Taking the scalar product of Eq. (105) by $\mathbf{u}$ gives

$$\frac{\partial}{\partial t}\Big(\frac{h}{2}|\mathbf{u}|^2\Big) + \nabla_0\cdot\Big(\mathbf{u}\big(\frac{h}{2}|\mathbf{u}|^2 + \frac{g}{2}h^2\big)\Big) + \mathbf{u}\cdot\nabla^{\gamma}_{sw}p - \frac{g}{2}h^2\nabla_0\cdot\mathbf{u} + gh\mathbf{u}\cdot\nabla_0 z_b = 0. \tag{113}$$

Using the duality relation (85) in (113) and adding Eq. (112) gives (110).

Besides, using Eq. (109) multiplied by $g/2$ and added to (110) gives (111). □

2.7 The Boundary Conditions

The set of equations (104)–(106) is completed with the following boundary conditions. We are considering a channel with an inlet Γ_{in} and an outlet Γ_{out} and we impose specific conditions on each of them, see Fig. 1. The inflow is imposed by a given discharge $\mathbf{q}_g$ on Γ_{in}, and a water depth h_g is imposed on Γ_{out}. Finally, we prescribe slip boundary conditions for the velocity at the walls of the channel Γ_s. Hence we have

$$h\mathbf{u}(t,\mathbf{x}) = \mathbf{q}_g(t,\mathbf{x}), \quad \text{on } \Gamma_{in}, \tag{114}$$

$$h(t,\mathbf{x}) = h_g(t,\mathbf{x}), \quad \text{on } \Gamma_{out}, \tag{115}$$

$$\mathbf{u}(t,\mathbf{x}) \cdot \mathbf{n} = 0, \quad \text{on } \Gamma_s. \tag{116}$$

Notice that we can replace the prescribed water depth at the outflow by a free outflow consisting in imposing a Neumann boundary condition over the elevation

$$\nabla_0 h \cdot \mathbf{n} = 0, \quad \text{on } \Gamma_{out}.$$

2.8 Dispersion Relation

The model (104)–(106) is a shallow water type model with compressible effects coming from the acoustic wave propagation. A fundamental question is to know what are the velocities of the waves propagating in such a model and typically the influence of the sound speed c over these velocites.

Let us consider the system (104)–(106) in the one-dimensional case, with flat topography and where the temperature variations are neglected. It has the form of an advection-reaction system, namely

$$\frac{\partial Y}{\partial t} + A\frac{\partial Y}{\partial x} + BY = 0, \tag{117}$$

with

$$Y = \begin{pmatrix} h \\ u \\ w \\ p \end{pmatrix}, \quad A = \begin{pmatrix} u & h & 0 & 0 \\ g + \frac{p}{h} & u & 0 & 1 \\ 0 & 0 & u & 0 \\ 0 & c^2 - \frac{gh}{2} & 0 & u \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{\gamma}{h} \\ 0 & 0 & \frac{\gamma c^2}{h} & 0 \end{pmatrix}.$$

Let us introduce $Y_0 \in \mathbb{R}^4$ and k, ω being two constants, namely the wave number and the frequency. A necessary condition so that the system (117) admits a solution having the form $Y = Y_0 e^{i(kx-\omega t)}$ is that

$$\det(i\omega I_4 - ikA + B) = 0,$$

I_4 being the identity matrix of dimension 4. This leads to the four roots

$$\frac{\omega}{k} = u \pm \frac{\sqrt{2}}{2}\sqrt{C_{sw,1} \pm \sqrt{(C_{sw,1})^2 - C_{sw,3}}}, \tag{118}$$

with $C_{sw,1} = c^2 + gh + p + c^2\frac{\gamma^2}{(hk)^2}$ and $C_{sw,2} = 4c^2\gamma^2(gh + p)$. As in (58), we can assume that $c \gg 1$ leading to the four approximated roots

$$u \pm \gamma\sqrt{\frac{gh + p}{\gamma^2 - (hk)^2}} + O\left(\frac{1}{c^2}\right), u \pm c\sqrt{1 + \frac{\gamma^2}{(hk)^2}} + O\left(\frac{1}{c}\right). \tag{119}$$

Remark 6 From the estimates (119), it appears that the model (104)–(106) is able to propagate both water waves and acoustic waves. But since we are in a shallow water context, we have $hk \ll 1$ and for the acoustic waves we do not exactly recover the expected velocities $u \pm c$.

Remark 7 In the context of wave propagation i.e. with flat bottom and assuming the water depth has the form $h = h_0 + f(kx - \omega t)$ with $h_0 = cst$ and $|f(.)| \ll h_0$, it is easy to see that the first term in (119) becomes

$$\gamma\sqrt{\frac{gh_0}{\gamma^2 - (hk)^2}} + O\left(\frac{1}{c^2}\right),$$

corresponding for $\gamma = \sqrt{3}$ in the context of large wavelength ($kh_0 \ll 1$) and up to $O((kh_0)^2)$ terms, to the classical Airy wave dispersion relation [1].

2.9 A Pseudo-compressible Model

As we have seen in the previous paragraph, if c is chosen corresponding to the sound speed in water, then the model (104)–(106) is able to propagate, in a shallow water context, both water and acoustic waves. But since $c \gg 1$, we introduce

$$\varepsilon = \frac{1}{c^2},$$

then the model (104)–(106) can be seen as a pseudo-compressible version of the model (9)–(11) allowing to derive an explicit in time numerical scheme that will be studied in the following section. More precisely Eq. (106) writes

$$\varepsilon\left(\frac{\partial}{\partial t}\big(hp+\frac{gh^2}{2}\big)+\nabla_0\cdot\Big(\big(hp+\frac{gh^2}{2}\big)\mathbf{u}\Big)\right)+\mathrm{div}_{sw}^{\gamma}(\mathbf{u})=0,$$

and corresponds to Eq. (11) when ε goes to 0. Following the results obtained in Proposition 6, it is important to notice that in the formulation (104)–(106), the limit $\varepsilon \to 0$ is not singular. Unlike the incompressible limit of compressible models, the limit when $\varepsilon \to 0$ of the model (104)–(106) is the model (9)–(11).

Hence the model (104)–(106) can be seen as

- a dispersive shallow water type model propagating water and acoustic waves,
- a pseudo-compressible dispersive model whose numerical resolution is easier to implement compared to a fully compressible model. This second aspect is studied in the two next sections.

Notice that several authors have proposed approximated versions of the divergence free constraint for dispersive models [19, 24], for which the origin of relaxation is not related to acoustics. The model formulation (104)–(106) is similar to the one studied in [20] but the derivation process—based on the so-called hyperbolic divergence cleaning [15]—differs. The numerical strategy proposed in [20] based on high order discontinuous Galerkin schemes is also different from the one presented hereafter.

3 The Numerical Scheme (Explicit in Time)

In this section, we propose and study a numerical scheme for the system (104)–(106) with $\varepsilon = 1/c^2$.

Let us introduce the notations

$$X=\begin{pmatrix} h\\ hu\\ hv\\ hw\end{pmatrix},\qquad F(X)=\begin{pmatrix} hu & hv\\ hu^2+\frac{g}{2}h^2 & huv\\ huv & hv^2+\frac{g}{2}h^2\\ huw & hvw\end{pmatrix},$$

and $S(X)=(0,-gh\nabla_0(z_b))^T$, $R=(0,\nabla_{sw}^{\gamma}p)^T$ where $\nabla_{sw}^{\gamma}p$ is defined by (12). Then, the system (104)–(106) can be written under the form

$$\frac{\partial X}{\partial t}+\nabla_{x,y}\cdot F(X)+R=S(X), \tag{120}$$

$$\varepsilon\Big(\frac{\partial}{\partial t}\big(hp+\frac{gh^2}{2}\big)+\nabla_0\cdot\Big(\big(hp+\frac{gh^2}{2}\big)\mathbf{u}\Big)\Big)+\mathrm{div}_{sw}^{\gamma}(\mathbf{u})=0. \tag{121}$$

3.1 Time Discretisation

Let us sketch the main steps of the procedure. We set t^0 the initial time and $t^{n+1} = t^n + \Delta t^n$ where Δt^n satisfies a stability condition (CFL) precised later—at the fully discrete level—and the state X^n denotes an approximation of $X(t^n)$. For each time step, we consider an intermediate state which will be denoted with the superscript $^{n+1/2}$. The first step consists in solving the Saint-Venant part of the system (120) with the topography source term and completed with the hyperbolic part of (121) in order to obtain the state $X^{n+1/2} = (h^{n+1/2}, (h\mathbf{u})^{n+1/2})^T$ and $(hp)^{n+1/2}$. Then the state X^{n+1} is computed taking into account the contribution of the non-hydrostatic pressure terms.

More precisely, in the system (120), (121) water waves generally propagate at a lower velocity than acoustic waves. Therefore, we propose an explicit time scheme—constrained by an associated CFL condition that will be precised in the fully discrete case, see (159)—for the Saint-Venant part of Eq. (120). For the dispersive terms, we adopt an iterative resolution scheme explicit in time and constrained by a generally more restrictive CFL condition associated with the sound speed. Hence, the proposed semi-discretization in time consists in the following time-splitting strategy

$$\begin{cases} X^{n+1/2} = X^n - \Delta t^n \nabla_{x,y} \cdot F(X^n) + \Delta t^n S(X^n), \\ (hp)^{n+1/2} = (hp)^n - \frac{g}{2}\Big(h^{n+1/2})^2 - h^n)^2\Big) \\ \qquad\qquad -\Delta t^n \nabla_0 \cdot \Big(h^n\big(p^n + \frac{g}{2}(h^n)^2\big)\mathbf{u}^n\Big), \end{cases} \tag{122}$$

$$\begin{cases} p^{n+1/2,k+1} = p^{n+1/2,k} - \frac{\Delta t^n}{\varepsilon K h^{n+1}} \mathrm{div}_{sw}^{\gamma} \mathbf{u}^{n+1/2,k}, \\ \mathbf{u}^{n+1/2,k+1} = \mathbf{u}^{n+1/2,k} - \frac{\Delta t^n}{K h^{n+1}} \nabla_{sw}^{\gamma} p^{n+1/2,k+1}, \end{cases} \quad k = 1, \ldots, K \tag{123}$$

with $p^{n+1/2,1} = p^{n+1/2}$, $\mathbf{u}^{n+1/2,1} = \mathbf{u}^{n+1/2}$, $p^{n+1} = p^{n+1/2,K+1}$, $\mathbf{u}^{n+1} = \mathbf{u}^{n+1/2,K+1}$ and where for the first component of X we have $h^{n+1} = h^{n+1/2}$ since the first component of R is zero. Notice that the two operators $\mathrm{div}_{sw}^{\gamma}$ and ∇_{sw}^{γ} are defined by Eqs. (12), (13) using h^{n+1}. K is an integer that is defined in order to ensure a stability condition for the acoustic-like wave propagation. The value of K is precised in the fully discrete case, see (162).

3.2 Influence of the Pseudo-compressibility over the Computational Costs

In [2], the authors have studied the model (1)–(5)—that is exactly the model (104)–(106) with $\varepsilon = 0$—and proposed the following semi-discretization in time

$$X^{n+1/2} = X^n - \Delta t^n \nabla_{x,y} \cdot F(X^n) + \Delta t^n S(X^n), \tag{124}$$

$$(h\mathbf{u})^{n+1} = (h\mathbf{u})^{n+1/2} - \Delta t^n \nabla_{sw}^{\gamma} p^{n+1}, \tag{125}$$

$$\mathrm{div}_{sw}^{\gamma} \mathbf{u}^{n+1} = 0, \tag{126}$$

in which Eq. (125) allows to correct the predicted value $X^{n+1/2}$ in order to obtain a state which satisfies the divergence free condition (126). The equation satisfied by the pressure is then an elliptic equation which is obtained by applying the shallow water divergence operator $\mathrm{div}_{sw}^{\gamma}$ to Eq. (125) and reads

$$\mathrm{div}_{sw}^{\gamma} \left(\frac{1}{h^{n+1}} \nabla_{sw}^{\gamma} p^{n+1} \right) = \frac{1}{\Delta t^n} \mathrm{div}_{sw}^{\gamma} \left(\frac{(h\mathbf{u})^{n+1/2}}{h^{n+1}} \right). \tag{127}$$

Once the pressure has been determined by the elliptic equation (127), the correction step (125) gives the final step X^{n+1}.

The main drawback of the time scheme (124)–(126) is the numerical cost of the resolution of Eq. (127). And Eq. (106) can be seen as a relaxed version of Eq. (126) allowing to replace the step (125)–(126) by the iterative method (123) applied to the model (120)–(121). More precisely, inserting the second equation of (123) (at iteration $k-1$) into the first one gives the relation

$$\begin{aligned} p^{k+1} &= p^k - \frac{\Delta t^n}{\varepsilon K} \mathrm{div}_{sw}^{\gamma} \mathbf{u}^{n+1/2,k-1} + \frac{(\Delta t^n)^2}{\varepsilon K^2 h^{n+1}} \mathrm{div}_{sw}^{\gamma} \left(\frac{1}{h^{n+1}} \nabla_{sw}^{\gamma} p^k \right), \\ &= 2p^k - p^{k-1} + \frac{(\Delta t^n)^2}{\varepsilon K^2 h^{n+1}} \mathrm{div}_{sw}^{\gamma} \left(\frac{1}{h^{n+1}} \nabla_{sw}^{\gamma} p^k \right), \end{aligned} \tag{128}$$

where the superscripts $^{n+1/2}$ have been dropped. Equation (128) appears as an explicit in time discretization of a wave equation. As expected, when ε tends to 0, Eq. (128) reduces to Eq. (127). Likewise, inserting the first equation of (123) into the second one gives the relation

$$\mathbf{u}^{k+1} = 2\mathbf{u}^k - \mathbf{u}^{k-1} + \frac{(\Delta t^n)^2}{\varepsilon K^2 h^{n+1}} \nabla_{sw}^{\gamma} \left(\frac{1}{h^{n+1}} \mathrm{div}_{sw}^{\gamma} \mathbf{u}^k \right). \tag{129}$$

The stability of the two discretizations (128), (129) will be examined in Sect. 4.4.

As already mentioned, if N is the number of cells in the considered mesh, the computational cost of the resolution of (127) is $O(N^{3/2})$ whereas the resolution of (123) is $O(KN) = O(N/\sqrt{\varepsilon})$. And hence, an estimation of ε is required to compare the costs of the explicit and implicit resolutions.

3.3 Choice of ε

If one is interested in the simulation of both water and acoustic waves, ε is chosen so that $\varepsilon = 1/c^2$, c being the sound speed. But if the objective is to approximate a relaxed version of the system (1)–(5) then ε is no more a physical parameter and has to be chosen so that the system (104)–(106) is a good approximation of the system (9)–(11). Hence, at each time step, ε can be chosen according to the computed values of the velocities and of the water depth.

And we can proceed as follows.

The energy of the model (104)–(106) behaves as

$$h\frac{u^2+v^2+w^2}{2}+\frac{g}{2}h\zeta+g\left(2-\frac{\gamma^2}{2}\right)hz_b+\frac{\varepsilon h}{2}\left(p+\frac{g}{2}h\right)^2.$$

Hence, we have to choose ε such that

$$\varepsilon\left(p+\frac{g}{2}h\right)^2 \ll u^2+v^2+w^2+g(\eta+z_b)+g\left(4-\gamma^2\right)z_b. \tag{130}$$

Another possibility is to recall that ε is related to the sound speed with $\varepsilon = 1/c^2$ and hence ε has to satisfy

$$\frac{1}{\sqrt{\varepsilon}} = c \gg |u|+|v|+\sqrt{gh},$$

i.e.

$$\varepsilon \ll \frac{1}{(|u|+|v|+\sqrt{gh})^2}. \tag{131}$$

The two conditions (130) and (131) are easy to implement and similar when $|u|+|v|+|w| \ll \sqrt{gh}$. But, in the context of dispersive flows (130) is more appropriate since the vertical velocity w is taken into account.

We have seen that the choice $\epsilon = 1/c^2$ corresponds to the propagation of acoustic waves. Smaller values of ε increase the computational costs of the scheme since it enlarges the value of the number of iterations K for the resolution of (123), see Sects. 3.1 and 3.2.

4 Detailed Numerical Scheme in 1d

A numerical scheme for the model (9)–(11) has been proposed and studied in [2]. Here we focus on the one dimensional case in order to prove the capability of the pseudo-compressible formulation.

In the one dimensional case, the model (104)–(106) writes

$$\frac{\partial h}{\partial t} + \frac{\partial (hu)}{\partial x} = 0, \tag{132}$$

$$\frac{\partial (hu)}{\partial t} + \frac{\partial}{\partial x}\left(hu^2 + \frac{g}{2}h^2 + hp\right) = -(gh + \frac{\gamma^2}{2}p)\frac{\partial z_b}{\partial x}, \tag{133}$$

$$\frac{\partial (hw)}{\partial t} + \frac{\partial (huw)}{\partial x} = \gamma p, \tag{134}$$

$$\varepsilon\left(\frac{\partial}{\partial t}\big(hp + \frac{gh^2}{2}\big) + \frac{\partial}{\partial x}\Big(\big(hp + \frac{gh^2}{2}\big)u\Big)\right) + \gamma w + h\frac{\partial u}{\partial x} - \frac{\gamma^2}{2}u\frac{\partial z_b}{\partial x} = 0. \tag{135}$$

In a more compact form and with obvious notations, the system (132)–(135) becomes

$$\frac{\partial X}{\partial t} + \frac{\partial F(X)}{\partial x} + R = S(X), \tag{136}$$

$$\varepsilon\Big(\frac{\partial (h\hat{p})}{\partial t} + \frac{\partial (hu\hat{p})}{\partial x}\Big) + \mathrm{div}^{\gamma}_{sw}(\mathbf{u}) = 0, \tag{137}$$

with $\mathbf{u} = (u, w)^T$, $\hat{p} = p + gh/2$ and

$$X = \begin{pmatrix} h \\ h\mathbf{u} \end{pmatrix}, \tag{138}$$

$$\nabla^{\gamma}_{sw} f = \begin{pmatrix} h\frac{\partial f}{\partial x} + \frac{\partial \zeta}{\partial x} f \\ -\gamma f \end{pmatrix}, \quad \mathrm{div}^{\gamma}_{sw}\mathbf{u} = \frac{\partial (hu)}{\partial x} - u\frac{\partial \zeta}{\partial x} + \gamma w. \tag{139}$$

Notice that the fundamental duality relation

$$\int_C p\ \mathrm{div}^{\gamma}_{sw}\mathbf{u}\, dx = [hup]_{\partial C} - \int_C \nabla^{\gamma}_{sw} p \cdot \mathbf{u}\, dx, \tag{140}$$

holds for any interval C.

The smooth solutions of Eqs. (132)–(135) satisfy the energy equality

$$\frac{\partial}{\partial t}\Big(\frac{h}{2}(u^2 + w^2) + \frac{g}{2}\left(2 - \frac{\gamma^2}{2}\right) + \frac{\varepsilon h}{2}\big(p + \frac{gh}{2}\big)^2\Big) + \frac{\partial}{\partial x}\Big(u(\frac{h}{2}(u^2 + w^2) + \frac{g}{2}\left(2 - \frac{\gamma^2}{2}\right) + \frac{\varepsilon h}{2}\big(p + \frac{gh}{2}\big)^2 + \frac{g}{2}h^2 + hp)\Big) = -\frac{\gamma g h}{2}w. \tag{141}$$

Introducing the 1d version of Eq. (109) given by

$$\frac{\partial (h\hat{\zeta})}{\partial t} + \frac{\partial}{\partial x}\big(h\hat{\zeta}u\big) = h\mathrm{div}^{\gamma}_{sw}(\mathbf{u}), \tag{142}$$

allows to have a conservative form of Eq. (141) under the form

$$\frac{\partial}{\partial t}\Big(\bar{E}+\frac{\varepsilon h}{2}\big(p+\frac{gh}{2}\big)^2\Big)+\frac{\partial}{\partial x}\Big(u\big(\bar{E}+\frac{\varepsilon h}{2}\big(p+\frac{gh}{2}\big)^2+\frac{g}{2}h^2+hp\big)\Big)=0, \tag{143}$$

with $\bar{E}=h(u^2+w^2)/2+gh(\eta+z_b)/2+g\frac{h\hat{\zeta}}{2}$, see Proposition 6.

4.1 Semi-discrete (in Time) Scheme

The 1d version of the time discretization (122)–(123) writes

$$\begin{cases} X^{n+1/2}=X^n-\Delta t^n\frac{\partial F(X^n)}{\partial x}+\Delta t^n S(X^n) \\ (h\hat{p})^{n+1/2}=(h\hat{p})^n-\Delta t^n\frac{\partial(h^n\hat{p}^n u^n)}{\partial x}\end{cases} \tag{144}$$

$$\begin{cases} p^{n+1/2,k+1}=p^{n+1/2,k}-\frac{\Delta t^n}{\varepsilon K h^{n+1}}\mathrm{div}^{\gamma}_{sw}\mathbf{u}^{n+1/2,k} \\ \mathbf{u}^{n+1/2,k+1}=\mathbf{u}^{n+1/2,k}-\frac{\Delta t^n}{Kh^{n+1}}\nabla^{\gamma}_{sw}p^{n+1/2,k+1}\end{cases} \tag{145}$$

with $p^{n+1/2,1}=p^{n+1/2}$, $\mathbf{u}^{n+1/2,1}=\mathbf{u}^{n+1/2}$ and $p^{n+1}=p^{n+1/2,K+1}$, $\mathbf{u}^{n+1}=\mathbf{u}^{n+1/2,K+1}$ where for the first component of X we have $h^{n+1}=h^{n+1/2}$.

The scheme (144)–(145) is explicit in time so it is important to examine its stability w.r.t. the discretisation step Δt^n, this will be done in Sect. 4.4.

4.2 The Semi-discrete (in Space) Scheme

To approximate the solution $X=(h,hu,hw)^T$, hp of the system (132)–(135), we use a combined finite volume/finite difference framework. We assume that the computational domain is discretized with I nodes x_i, $i=1,\dots,I$. We denote C_i the cell $(x_{i-1/2},x_{i+1/2})$ of length $\Delta x_i=x_{i+1/2}-x_{i-1/2}$ with $x_{i+1/2}=(x_i+x_{i+1})/2$. We denote $X_i=(h_i,q_{x,i},q_{z,i})^T$ with

$$X_i\approx\frac{1}{\Delta x_i}\int_{C_i}X(t,x)dx,$$

the approximate solution at time t on the cell C_i with $q_{x,i}=h_iu_i$, $q_{z,i}=h_iw_i$. Likewise, for the topography, we define

$$z_{b,i}=\frac{1}{\Delta x_i}\int_{C_i}z_b(x)dx.$$

The non-hydrostatic part of the pressure is discretized on a staggered grid

$$p_{i+1/2}\approx\frac{1}{\Delta x_{i+1/2}}\int_{x_i}^{x_{i+1}}p(t,x)dx,$$

with $\Delta x_{i+1/2} = x_{i+1} - x_i$ and we set $\hat{p}_{i+1/2} = p_{i+1/2} + g h_{i+1/2}/2$ where $h_{i+1/2}$ is defined by $\Delta x_{i+1/2} h_{i+1/2} = (\Delta x_i h_i + \Delta x_{i+1} h_{i+1})/2$.

Now we propose and study the semi-discrete (in space) scheme approximating the model (136)–(137). The semi-discrete scheme writes

$$\Delta x_i \frac{\partial X_i}{\partial t} + \left(F_{i+1/2-} - F_{i-1/2+}\right) + R_i = 0, \tag{146}$$

$$\Delta x_{i+1/2} \varepsilon \frac{\partial}{\partial t}\left(h_{i+1/2}\hat{p}_{i+1/2}\right) + \varepsilon\left(F_{\hat{p},i+1} - F_{\hat{p},i}\right) + \mathrm{div}^{\gamma}_{sw,i+1/2}\left(\{\mathbf{u}_j\}\right) = 0, \tag{147}$$

with the numerical fluxes

$$F_{i+1/2+} = \mathcal{F}(X_i, X_{i+1}, z_{b,i}, z_{b,i+1}) + \mathcal{S}_{i+1/2+} \tag{148}$$

$$F_{i+1/2-} = \mathcal{F}(X_i, X_{i+1}, z_{b,i}, z_{b,i+1}) + \mathcal{S}_{i+1/2-}. \tag{149}$$

$\mathcal{F}$ is a numerical flux for the conservative part of the system, $\mathcal{S}$ is a convenient discretization of the topography source term.

Since the first two lines of (136) correspond to the classical Saint-Venant system, the numerical fluxes

$$F_{i+1/2\pm} = \begin{pmatrix} F_{h,i+1/2} \\ F_{q_x,i+1/2\pm} \\ F_{q_z,i+1/2} \end{pmatrix}, \tag{150}$$

can be constructed using any numerical solver for the Saint-Venant system. More precisely for $F_{h,i+1/2}$,$F_{q_x,i+1/2\pm}$ we adopt numerical fluxes suitable for the Saint-Venant system with topography [10, 22, 25]. Notice that from the definition (138), since only the second component of $S(X)$ is non zero, only F_{q_x} has two interface values under the form $F_{q_x,i+1/2\pm}$. For the definition of $F_{q_z,i+1/2}$, the formula (see [5])

$$F_{q_z,i+1/2} = F_{h,i+1/2} w_{i+1/2}, \tag{151}$$

with

$$w_{i+1/2} = \begin{cases} w_i & \text{if } F_{h,i+1/2} \geq 0 \\ w_{i+1} & \text{if } F_{h,i+1/2} < 0 \end{cases} \tag{152}$$

can be used. The fluxes $F_{\hat{p},i}$ are defined similarly to (151), (152) but on the staggered grid by the following formula

$$F_{\hat{p},i} = \frac{F_{h,i+1/2} + F_{h,i-1/2}}{2} \hat{p}_i, \tag{153}$$

with

$$\hat{p}_i = \begin{cases} \hat{p}_{i-1/2} & \text{if } \frac{F_{h,i+1/2}+F_{h,i-1/2}}{2} \geq 0 \\ \hat{p}_{i+1/2} & \text{if } \frac{F_{h,i+1/2}+F_{h,i-1/2}}{2} < 0. \end{cases}$$

Combining the finite volume approach for the hyperbolic part with a finite difference strategy for the parabolic part, the non-hydrostatic part R_i is defined by

$$R_i = \begin{pmatrix} 0 \\ \nabla_{sw,i}^{\gamma}\, p \end{pmatrix},$$

where the two components of $\nabla_{sw,i}^{\gamma}\, p$ are defined (see (139)) by

$$\Delta x_i\ \nabla_{sw,i}^{\gamma}\, p\big|_1 = h_i(p_{i+1/2} - p_{i-1/2}) + \frac{p_{i+1/2}}{2}\big(\zeta_{i+1} - \zeta_i\big) + \frac{p_{i-1/2}}{2}\big(\zeta_i - \zeta_{i-1}\big), \tag{154}$$

$$\Delta x_i\ \nabla_{sw,i}^{\gamma}\, p\big|_2 = -\frac{\gamma}{2}\Big(\Delta x_{i+1/2} p_{i+1/2} + \Delta x_{i-1/2} p_{i-1/2}\Big), \tag{155}$$

with $\zeta_i = h_i + \frac{\gamma^2}{2} z_{b,i}$. And in (147), $\mathrm{div}_{sw,i+1/2}^{\gamma}(\mathbf{u})$ is defined by

$$\begin{aligned}\Delta x_{i+1/2}\mathrm{div}_{sw,i+1/2}^{\gamma}(\mathbf{u}) =& \frac{h_{i+1}+h_i}{2}(u_{i+1}-u_i) - \frac{u_i+u_{i+1}}{2}\big(z_{b,i+1}-z_{b,i}\big) \\ &+ \frac{\gamma\,\Delta x_{i+1/2}}{2}\big(w_{i+1}+w_i\big) \\ =& (hu)_{i+1} - (hu)_i - \frac{u_i+u_{i+1}}{2}\big(\zeta_{i+1}-\zeta_i\big) \\ &+ \frac{\gamma\,\Delta x_{i+1/2}}{2}\big(w_{i+1}+w_i\big).\end{aligned} \tag{156}$$

Notice that in the definitions (154)–(155) and in the sequel, the quantity p means $\{p_j\}$. Likewise in Eq. (156) and in the sequel, $\mathbf{u}$ means $\{\mathbf{u}_j\}$ for $1 \le j \le I$.

4.3 Wet-Dry Interface

The method presented above supposes that the water depth does not vanish since the resolution of (145) requires dividing the shallow water gradient and divergence operators by h. We use a strategy similar to [2, paragraph 5.2] that can be viewed as a Dirichlet condition on the dry zone of the domain, such that the non-hydrostatic pressure is solved only on the wet domain.

In practice, we introduce a small parameter $\theta \ll 1$ and the definitions (154)–(155) become

$$\begin{aligned}
\Delta x_i \ \nabla^{\gamma}_{sw,i} p\Big|_1 &= \mathbf{1}_{h_i \geq h_\theta}\Big(h_i(p_{i+1/2} - p_{i-1/2}) + \frac{p_{i+1/2}}{2}(\zeta_{i+1} - \zeta_i) + \frac{p_{i-1/2}}{2}(\zeta_i - \zeta_{i-1})\Big),\\
\Delta x_i \ \nabla^{\gamma}_{sw,i} p\Big|_2 &= -\frac{\gamma \mathbf{1}_{h_i \geq h_\theta}}{2}\Big(\Delta x_{i+1/2} p_{i+1/2} + \Delta x_{i-1/2} p_{i-1/2}\Big),\\
\Delta x_{i+1/2} \mathrm{div}^{\gamma}_{sw,i+1/2}(\mathbf{u}) &= \mathbf{1}_{h_i \geq h_\theta}\Big((hu)_{i+1} - (hu)_i - \frac{u_i + u_{i+1}}{2}(\zeta_{i+1} - \zeta_i)\\
&\quad + \frac{\gamma \Delta x_{i+1/2}}{2}(w_{i+1} + w_i)\Big),
\end{aligned}$$

with $h_\theta = \max(h, \theta)$.

4.4 Stability of the Scheme

Using the definitions (144), (145), (146), (147), (154) and (155), the fully discrete scheme for the system (136)–(137) writes

$$\begin{cases} X_i^{n+1/2} = X_i^n - \frac{\Delta t^n}{\Delta x_i}\left(F^n_{i+1/2-} - F^n_{i-1/2+}\right), \\ (h\hat{p})^{n+1/2}_{i+1/2} = (h\hat{p})^n_{i+1/2} - \frac{\Delta t^n}{\Delta x_{i+1/2}}\left(F^n_{\hat{p},i+1} - F^n_{\hat{p},i}\right), \end{cases} \tag{157}$$

$$\begin{cases} p^{n+1/2,k+1}_{i+1/2} = p^{n+1/2,k}_{i+1/2} - \frac{\Delta t^n}{\varepsilon K h^{n+1}} \mathrm{div}^{\gamma}_{sw,i+1/2} \mathbf{u}^{n+1/2,k}, \\ \mathbf{u}_i^{n+1/2,k+1} = \mathbf{u}_i^{n+1/2,k} - \frac{\Delta t^n}{K h^{n+1}} \nabla^{\gamma}_{sw,i} p^{n+1/2,k+1}. \end{cases} \tag{158}$$

The first equation of (157) gives a finite volume scheme for the Saint-Venant system. The choice of numerical fluxes $F_{i+1/2\pm}$ (see [10]) coupled with a numerical treatment of the topography source term e.g. using the hydrostatic reconstruction [3] gives a numerical resolution of the Saint-Venant system endowed with strong stability properties [4] that are recalled in Propositions 7 and 8. In Eqs. (157)–(158), Δt^n satisfies a CFL condition having the form

$$\Delta t^n = \max_{i \in I} \frac{\Delta x_i}{|V_i^n|}, \tag{159}$$

where V_i^n is related to the eigenvalues of the Saint-Venant system, see [10]. Since the expression of the numerical fluxes (Rusanov, HLL, kinetic solver…) is not precised we are not able to give the exact expression of the CFL condition. In order to study the discrete energy balance induced by the numerical scheme (157)–(158), we define a discrete version of (142) under the form

$$\begin{aligned}
\frac{\Delta t^n}{\Delta x_i}(h\hat{\zeta})_i^{n+1/2,k+1} &= \frac{\Delta x_i}{\Delta t^n}(h\hat{\zeta})_i^{n+1/2,k} - \frac{1}{K}\left(\hat{\zeta}^{n+1/2,k}_{i+1/2} F_{h,i+1/2} - \hat{\zeta}^{n+1/2,k}_{i-1/2} F_{h,i-1/2}\right)\\
&+ \frac{\Delta x_{i+1/2} h^{n+1}_{i+1/2}}{2K} \mathrm{div}^{\gamma}_{sw,i+1/2}(\mathbf{u}^{n+1/2,k}) + \frac{\Delta x_{i+1/2} h^{n+1}_{i-1/2}}{2K} \mathrm{div}^{\gamma}_{sw,i-1/2}(\mathbf{u}^{n+1/2,k}),
\end{aligned} \tag{160}$$

where $\hat{\zeta}_{i+1/2}$ is defined by

$$\hat{\zeta}_{i+1/2} = \begin{cases} \hat{\zeta}_i & \text{if } F_{h,i+1/2} \geq 0 \\ \hat{\zeta}_{i+1} & \text{if } F_{h,i+1/2} < 0. \end{cases}$$

Now we focus on the stability condition for the resolution of (145) or equivalently (128). Using the definitions (154) and (155), we obtain the discrete version of the operator

$$\Delta_{sw}^{\gamma} p = \mathrm{div}_{sw}^{\gamma}\Big(\frac{1}{h}\nabla_{sw}^{\gamma} p\Big),$$

with $D_{i+1/2}p = -\Delta x_{i+1/2}\Delta_{sw,i+1/2}^{\gamma} p$ and

$$\begin{aligned} D_{i+1/2}p = &- \frac{h_{i+1}}{\Delta x_{i+1}}\Big(p_{i+3/2} - p_{i+1/2}\Big) + \frac{h_i}{\Delta x_i}\Big(p_{i+1/2} - p_{i-1/2}\Big) \\ &- \frac{p_{i+3/2}}{2\Delta x_{i+1}}\Big(\zeta_{i+2} - 2\zeta_{i+1} + \zeta_i\Big) - \frac{\Delta x_i - \Delta x_{i+1}}{\Delta x_{i+1}\Delta x_i} p_{i+1/2}\Big(\zeta_{i+1} - \zeta_i\Big) \\ &- \frac{p_{i-1/2}}{2\Delta x_i}\Big(\zeta_{i+1} - 2\zeta_i + \zeta_{i-1}\Big) \\ &+ \frac{p_{i+3/2}}{4h_{i+1}\Delta x_{i+1}}\Big(\zeta_{i+2} - \zeta_{i+1}\Big)\Big(\zeta_{i+1} - \zeta_i\Big) \\ &+ \frac{p_{i+1/2}}{4}\left(\frac{1}{h_{i+1}\Delta x_{i+1}} + \frac{1}{h_i\Delta x_i}\right)\Big(\zeta_{i+1} - \zeta_i\Big)^2 \\ &+ \frac{p_{i-1/2}}{4h_i\Delta x_i}\Big(\zeta_{i+1} - \zeta_i\Big)\Big(\zeta_i - \zeta_{i-1}\Big) \\ &+ \frac{\gamma^2\Delta x_{i+1/2}}{4}\left(\frac{\Delta x_{i+3/2}p_{i+3/2} + \Delta x_{i+1/2}p_{i+1/2}}{\Delta x_{i+1}h_{i+1}}\right. \\ &+ \left.\frac{\Delta x_{i+1/2}p_{i+1/2} + \Delta x_{i-1/2}p_{i-1/2}}{\Delta x_i h_i}\right). \end{aligned} \tag{161}$$

Using the expression (161), we are now able to precise the CFL type stability condition for the discretized version of Eq. (128) that writes

$$\begin{aligned} 2 - \frac{(\Delta t^n)^2}{\varepsilon K^2 h_{i+1/2}\Delta x_{i+1/2}}&\left(\frac{h_{i+1}}{\Delta x_{i+1}} + \frac{h_i}{\Delta x_i} + \frac{1}{4}\left(\frac{1}{h_{i+1}\Delta x_{i+1}} + \frac{1}{h_i\Delta x_i}\right)\Big(\zeta_{i+1} - \zeta_i\Big)^2\right. \\ &\left. - \frac{\Delta x_i - \Delta x_{i+1}}{\Delta x_{i+1}\Delta x_i}\Big(\zeta_{i+1} - \zeta_i\Big) + \frac{\gamma^2\Delta x_{i+1/2}^2}{4}\left(\frac{1}{h_{i+1}\Delta x_{i+1}} + \frac{1}{h_i\Delta x_i}\right)\right) \geq 0, \end{aligned}$$

that is fulfilled for

$$K^2 \geq \frac{(\Delta t^n)^2}{2\varepsilon h_{i+1/2}\Delta x_{i+1/2}}\left(\frac{h_{i+1}}{\Delta x_{i+1}}+\frac{h_i}{\Delta x_i}+\frac{1}{4}\left(\frac{1}{h_{i+1}\Delta x_{i+1}}+\frac{1}{h_i\Delta x_i}\right)\left(\zeta_{i+1}-\zeta_i\right)^2 + \frac{|\Delta x_i-\Delta x_{i+1}|}{\Delta x_{i+1}\Delta x_i}|\zeta_{i+1}-\zeta_i|+\frac{\gamma\Delta x_{i+1/2}^2}{4}\left(\frac{1}{h_{i+1}\Delta x_{i+1}}+\frac{1}{h_i\Delta x_i}\right)\right). \quad (162)$$

And the condition (162) is satisfied when

$$K^2 \geq \frac{(\Delta t^n)^2}{2\varepsilon h_{min}\Delta x_{min}^2}\left(2h_{max}+\frac{2}{h_{min}}(\delta\zeta)_{max}^2+(\delta\zeta)_{max}+\frac{\gamma^2\Delta x_{max}^2}{2h_{min}}\right),$$

with $r_{min}=\min_{1\leq i\leq I} r_i$, $r_{max}=\max_{1\leq i\leq I} r_i$ for $r=h,\Delta x$ and $\delta\zeta_{max}=\max_{1\leq i\leq I-1}|\zeta_{i+1}-\zeta_i|$.

The fully discrete scheme (157), (158) satisfies the following stability properties.

Proposition 7 *Assuming a suitable CFL condition* (159) *adpated to the chosen numerical fluxes* (150) *for the hyperbolic part, the scheme obtained by coupling the semi-discretizations* (144), (145) *and* (146), (147)

(i) preserves the nonnegativity of the water depth i.e. $h_i^n \geq 0$, $\forall i$, $\forall n$,
(ii) preserves the steady state of the lake at rest,
(iii) is consistent with the model (136)–(137).

Let us consider that, under a suitable CFL condition associated with the time discretization (144) and the chosen numerical fluxes $F_{h,i\pm1/2}$ and $F_{q_x,i\pm1/2}$ in (150), the numerical approximation of the Saint-Venant part of Eq. (136) allows to obtain a discrete entropy equality under the form

$$\Delta x_i\left(E_i^{sv}\right)^{n+1/2}=\Delta x_i\left(E_i^{sv}\right)^n-\Delta t^n\left(\mathcal{G}_{i+1/2}^n-\mathcal{G}_{i-1/2}^n\right)+\mathcal{D}_i^n, \quad (163)$$

with $E_i^{sv}=\frac{h_i}{2}u_i^2+\frac{g}{2}(\eta_i^2-z_{b,i}^2)$ and where $\mathcal{G}_{i\pm1/2}^n$ are numerical fluxes. $\mathcal{D}_i^n$ is a nonpositive term and contains typically two different contributions: the numerical dissipation coming from the upwinding in the space discretization and the error due to the explicit time scheme.

Then assuming (163), we now prove that the numerical scheme (157), (158) satisfies a discrete entropy equality.

Proposition 8 *Assuming* (163), *the scheme* (144), (145), (146),(147) *satisfies the following discrete entropy equality*

$$\Delta x_i\bar{E}_i^{n+1}=\Delta x_i\bar{E}_i^n-\Delta t^n\left(\bar{\mathcal{G}}_{i+1/2}^n-\bar{\mathcal{G}}_{i-1/2}^n\right)+\bar{\mathcal{D}}_i^n, \quad (164)$$

where $\bar{E}_i=E_i^{sv}+\frac{h_i}{2}w_i^2+g(h\hat{\zeta})_i+\frac{\varepsilon}{2}\widetilde{h_i\hat{p}_i^2}$ *and*

$$\bar{\mathcal{G}}^n_{i+1/2} = \mathcal{G}^n_{i+1/2} + F^n_{h,i+1/2}\frac{(w^n_{i+1/2})^2}{2} + \varepsilon F^n_{h,i+1}\frac{(\hat{p}^n_{i+1})^2}{2} + \frac{1}{K}\sum_{k=1}^{K}\big(h^{n+1}_{i+1/2}u^{n+1/2,k}_{i+1/2}p^{n+1/2,k+1}_{i+1/2} + \hat{\xi}^{n+1/2,k}_{i+1/2}F_{h,i+1/2}\big),$$

$$\Delta x_i\widetilde{(h_i\hat{p}^2_i)}^{n+1} = \frac{1}{2K}\sum_{k=1}^{K}\left(\Delta x_{i+1/2}\frac{h^{n+1}_{i+1/2}}{2}\left(\hat{p}^{n+1/2,k+1}_{i+1/2}\right)^2 + \Delta x_{i-1/2}\frac{h^{n+1}_{i-1/2}}{2}\left(\hat{p}^{n+1/2,k+1}_{i-1/2}\right)^2\right),$$

with

$$\begin{aligned}
\Delta x_i\bar{\mathcal{D}}^n_i = {} & \Delta x_i\mathcal{D}^n_i + \Delta t^n\left(\left[F^n_{h,i+1/2}\right]_-(w^n_{i+1}-w^n_i)^2 - \left[F^n_{h,i-1/2}\right]_+(w^n_i-w^n_{i-1})^2\right)\\
& + \frac{\Delta t^n}{2}\left(\frac{F^n_{h,i+1}}{2}(\hat{p}^n_{i+1}-\hat{p}^n_{i+1/2})^2 - \frac{F^n_{h,i}}{2}(\hat{p}^n_i-\hat{p}^n_{i+1/2})^2\right)\\
& + \frac{\Delta t^n}{2}\left(\frac{F^n_{h,i}}{2}(\hat{p}^n_i-\hat{p}^n_{i-1/2})^2 - \frac{F^n_{h,i-1}}{2}(\hat{p}^n_{i-1}-\hat{p}^n_{i-1/2})^2\right)\\
& + \frac{\Delta x_i h^{n+1}_i}{2}(w^{n+1/2}_i - w^n_i)^2 + \frac{\Delta x_{i+1/2}h^{n+1}_{i+1/2}}{2}(\hat{p}^{n+1/2}_{i+1/2}-\hat{p}^n_{i+1/2})^2\\
& + \frac{\Delta x_{i-1/2}h^{n+1}_{i-1/2}}{2}(\hat{p}^{n+1/2}_{i-1/2}-\hat{p}^n_{i-1/2})^2\\
& - \sum_{k=1}^{K}\frac{\Delta x_{i+1/2}h^{n+1}_{i+1/2}}{2}\left(\varepsilon\left(\hat{p}^{n+1/2,k+1}_{i+1/2}-\hat{p}^{n+1/2,k}_{i+1/2}\right)^2 - \left|\mathbf{u}^{n+1/2,k+1}_{i+1/2}-\mathbf{u}^{n+1/2,k}_{i+1/2}\right|^2\right)\\
& - \sum_{k=1}^{K}\frac{\Delta x_{i+1/2}h^{n+1}_{i-1/2}}{2}\left(\varepsilon\left(\hat{p}^{n+1/2,k+1}_{i-1/2}-\hat{p}^{n+1/2,k}_{i-1/2}\right)^2 - \left|\mathbf{u}^{n+1/2,k+1}_{i-1/2}-\mathbf{u}^{n+1/2,k}_{i-1/2}\right|^2\right).
\end{aligned}$$

Remark 8 When considering the semi-descrete in space scheme detailed in Sect. 4.2, a semi-discrete in space version of (164) holds where all the non-negative terms in the expression of $\bar{\mathcal{D}}^n_i$ corresponding to time discretisation errors vanish.

Proof ***(Proposition*** 7***)*** *(i)* The statement that $\mathcal{F}$ preserves the nonnegativity of the water depth means exactly that

$$F_h(h_i = 0, u_i, h_{i+1}, u_{i+1}) - F_h(h_{i-1}, u_{i-1}, h_i = 0, u_i) \le 0,$$

for all choices of the other arguments. From (144), (146), (148) and (149), we need to check that, with obvious notations

$$F_h(X^n_{i+1/2-}, X^n_{i+1/2+}) - F_h(X^n_{i-1/2-}, X^n_{i-1/2+}) \le 0,$$

whenever $h^n_i = 0$. And this property holds typically when the hydrostatic reconstruction (HR) is used to approximate the topography source term since for the HR technique $h_i = 0$ implies $h_{i+1/2-} = h_{i-1/2+} = 0$, see [3].

(ii) When $u^n_i = 0$ for all i, the properties of the hydrostatic reconstruction technique ensure $F^n_{i+1/2-} = F^n_{i-1/2+}$ in (144), (146) and $F^n_{p,i+1-} = F^n_{p,i+}$ in (147). More-

over since $u_i^n = 0\ \forall i$ we have $R_i = 0$ in (146) and $\mathrm{div}_{sw,i+1/2}^{\gamma}(\{\mathbf{u}\}) = 0$ in (147). Therefore $\forall i$

$$X_i^{n+1} = X_i^n, \quad \text{and} \quad p_{i+1/2}^{n+1} = p_{i+1/2}^n,$$

proving that the scheme is well-balanced.

(iii) The discretization (144), (145) is an explicit first order time scheme. The numerical fluxes defined by (148), (149) and (153) are a consistent discretization of the hyperbolic part of the system (136), (137) without topography. Likewise, the hydrostatic reconstruction applied to the fluxes (150), (153) gives a consistent discretization of the system (136), (137) with topography and the discretizations (154), (155) being obviously consistent with the dispersive part, this proves the result. □

Proof ***(Proposition 8)*** Since we have assumed that the kinetic energy of the Saint-Venant part of Eq. (136) satisfies (163), this means that the first two components of the first equation of (157) multiplied respectively by $gh_i^n - (u_i^n)^2/2$ and u_i^n give Eq. (163). It remains to consider the contributions to the energy balance of the last two components of (157) and of Eq. (158).

First let us multiply the third component of the first equation of (157) by w_i^n, then we get

$$\begin{aligned}
&\frac{h_i^{n+1}}{2}(w_i^{n+1/2})^2 - \frac{h_i^n}{2}(w_i^n)^2 + \frac{\Delta t^n}{\Delta x_i}\left(F_{h,i+1/2}^n \frac{(w_{i+1/2}^n)^2}{2} - F_{h,i-1/2}^n \frac{(w_{i-1/2}^n)^2}{2}\right) \\
= \ &\frac{\Delta t^n}{\Delta x_i}\left(\left[F_{h,i+1/2}^n\right]_- (w_{i+1}^n - w_i^n)^2 - \left[F_{h,i-1/2}^n\right]_+ (w_i^n - w_{i-1}^n)^2\right) + \frac{h_i^{n+1}}{2}(w_i^{n+1/2} - w_i^n)^2,
\end{aligned}$$

with the notations $[a]_+ = max(a,0), [a]_- = min(a,0)\ a = [a]_+ + [a]_-$ and $w_{i+1/2}^n$ is defined by (152). Then we multiply the last component of Eq. (157) by $p_{i+1/2}^n + \frac{g}{2}h_{i+1/2}^n$ leading to

$$\begin{aligned}
&\frac{h_{i+1/2}^{n+1}}{2}\left(\hat{p}_{i+1/2}^{n+1/2}\right)^2 - \frac{h_{i+1/2}^n}{2}\left(\hat{p}_{i+1/2}^n\right)^2 + \frac{\Delta t^n}{\Delta x_{i+1/2}}\left(F_{h,i+1}^n \frac{(\hat{p}_{i+1}^n)^2}{2} - F_{h,i}^n \frac{(\hat{p}_i^n)^2}{2}\right) \\
&= \frac{\Delta t^n}{\Delta x_{i+1/2}}\left(\frac{F_{h,i+1}^n}{2}(\hat{p}_{i+1}^n - \hat{p}_{i+1/2}^n)^2 - \frac{F_{h,i}^n}{2}(\hat{p}_i^n - \hat{p}_{i+1/2}^n)^2\right) \\
&\quad + \frac{h_{i+1/2}^{n+1}}{2}\left(\hat{p}_{i+1/2}^{n+1/2} - \hat{p}_{i+1/2}^n\right)^2,
\end{aligned}$$

with $F_{h,i+1} = (F_{h,i+3/2} + F_{h,i+1/2})/2$. Thanks to the definition (151), the two quantities

$$\frac{F_{h,i+1}^n}{2}(\hat{p}_{i+1}^n - \hat{p}_{i+1/2}^n)^2, \quad \text{and} \quad -\frac{F_{h,i}^n}{2}(\hat{p}_i^n - \hat{p}_{i-1/2}^n)^2,$$

are always non-positive.

Second, we multiply the equations (158) respectively by $\hat{p}_{i+1/2}^{n+1/2,k+1}$ and $\mathbf{u}_i^{n+1/2,k}$ and sum the obtained relations for $k = 1, \ldots K$. Precisely, starting from the definitions (154), (155), we rewrite $\nabla_{sw,i}^{\gamma} p$ under the form

$$\nabla^{\gamma}_{sw,i}\, p = \nabla^{\gamma}_{sw,i+1/2-} p + \nabla^{\gamma}_{sw,i-1/2+} p,$$

with

$$\nabla^{\gamma}_{sw,i+1/2-} p = \left| \begin{array}{l} h_i (p_{i+1/2} - p_i) + \frac{p_{i+1/2}}{2} \big(\zeta_{i+1} - \zeta_i\big) \\ \qquad - \frac{\gamma}{2} \Delta x_{i+1/2}\, p_{i+1/2} \end{array} \right.$$

and we obtain a discrete version of the duality relation (140) under the form

$$\begin{aligned} \Delta x_i \nabla^{\gamma}_{sw,i+1/2-} p^{n+1/2,k+1} \cdot \mathbf{u}_i^{n+1/2,k} &= e^{n+1/2,k+1/2}_{i+1/2-} \\ &\quad - \frac{\Delta x_{i+1/2}}{2} \mathrm{div}^{\gamma}_{sw,i+1/2} \big(\mathbf{u}^{n+1/2,k}\big)\, p^{n+1/2,k+1}_{i+1/2}, \end{aligned} \tag{165}$$

with $e^{n+1/2,k+1/2}_{i+1/2-}$ defined by

$$\begin{aligned} e^{n+1/2,k+1/2}_{i+1/2-} &= h^{n+1}_{i+1/2} u^{n+1/2,k}_{i+1/2}\, p^{n+1/2,k+1}_{i+1/2} - h^{n+1}_i u^{n+1/2,k}_i\, p^{n+1/2,k+1}_i \\ &\quad + p^{n+1/2,k+1}_{i+1/2} \big(\zeta_{i+1} - \zeta_i\big) \frac{u^{n+1/2,k}_i - u^{n+1/2,k}_{i+1}}{2} \\ &\quad - \frac{\gamma}{4} \Delta x_{i+1/2}\, p^{n+1/2,k+1}_{i+1/2} (w^{n+1/2,k}_i - w^{n+1/2,k}_{i+1}). \end{aligned}$$

Notice that in the expression of $e^{n+1/2,k+1/2}_{i+1/2-}$ the last two terms are second order and $e^{n+1/2,k+1/2}_{i+1/2+} + e^{n+1/2,k+1/2}_{i+1/2-} = h^{n+1}_{i+1} u^{n+1/2,k}_{i+1}\, p^{n+1/2,k+1}_{i+1} - h^{n+1}_i u^{n+1/2,k}_i\, p^{n+1/2,k+1}_i$.

The duality relation (165) has been written for the variable $p^{n+1/2,k+1}_{i\pm1/2}$ but the last two terms in (165) should be a discrete version of the r.h.s. of Eq. (112) i.e. of the quantity $\hat{p}\mathrm{div}^{\gamma}_{sw}(\mathbf{u})$. And since $\hat{p} = p + gh/2$ the reminder is (for interfaces $i \pm 1/2$)

$$\frac{g}{2} h^{n+1}_{i+1/2} \mathrm{div}^{\gamma}_{sw,i+1/2} \big(\mathbf{u}^{n+1/2,k}\big) + \frac{g}{2} h^{n+1}_{i-1/2} \mathrm{div}^{\gamma}_{sw,i-1/2} \big(\mathbf{u}^{n+1/2,k}\big),$$

corresponding to the right hand side of (160) multiplied by g.

For the errors coming from the time dicretization of Eqs. (158), we have

$$\begin{aligned} \Big((hp)^{n+1/2,k+1}_{i+1/2} - (hp)^{n+1/2,k}_{i+1/2}\Big) \hat{p}^{n+1/2,k+1}_{i+1/2} &= \Big((h\hat{p})^{n+1/2,k+1}_{i+1/2} - (h\hat{p})^{n+1/2,k}_{i+1/2}\Big) \hat{p}^{n+1/2,k+1}_{i+1/2} \\ &= \frac{1}{2} (h\hat{p}^2)^{n+1/2,k+1}_{i+1/2} - \frac{1}{2} (h\hat{p}^2)^{n+1/2,k}_{i+1/2} \\ &\quad + \frac{h^{n+1}_{i+1/2}}{2} \Big(\hat{p}^{n+1/2,k+1}_{i+1/2} - \hat{p}^{n+1/2,k}_{i+1/2}\Big)^2, \\ \Big((h\mathbf{u})^{n+1/2,k+1}_i - (h\mathbf{u})^{n+1/2,k}_i\Big) \cdot \mathbf{u}^{n+1/2,k}_i &= \frac{h^{n+1}_i}{2} |\mathbf{u}^{n+1/2,k+1}_i|^2 - \frac{h^{n+1}_i}{2} |\mathbf{u}^{n+1/2,k}_i|^2 \\ &\quad - \frac{h^{n+1}_{i+1/2}}{2} \Big|\mathbf{u}^{n+1/2,k+1}_{i+1/2} - \mathbf{u}^{n+1/2,k}_{i+1/2}\Big|^2. \end{aligned}$$

Summing the previous relations for $k = 1, \ldots, K$ and adding the result to the other contributions gives the corresponding expressions appearing in relation (164). This ends the proof. □

Remark 9 For the discretization of the model (132)–(135), we have presented a first order scheme in space and time. Second order extensions (in space and time) can be proposed, following [2].

4.5 Simulation Results

In this paragraph, only few numerical examples are presented. A more complete validation of the numerical procedure will be presented in a companion paper. Notice that in the 1d case, we mainly validate the numerical scheme but the reduction of the computation costs will be more significant in a two-dimensional setting.

4.5.1 Dingemans Experiments

The experiments carried out by Dingemans [18] at Delft Hydraulics deal with the wave propagation over uneven bottoms. A wave generator produces a small amplitude wave (0.02 m) at the left boundary of a basin with vertical shores. A vertical shore closes the basin at the right boundary, due to the considered time window, the measurements are not perturbed by the reflected wave on the right boundary. At rest, the water depth in the channel varies from 0.4 m to 0.1 m, see Fig. 3. Eight sensors recording the free surface elevation are located at abscissa 2 m, 4 m, 10.5 m, 12.5 m, 13.5 m, 14.5 m, 15.7 m and 17.3 m.

For $\gamma = \sqrt{3}$, we compare the simulation results obtained with the two numerical schemes (the one proposed in [2] and the one proposed in this paper with $\varepsilon = 1/c^2 = 10^{-4}$ m^{-2}.s^2). The results obtained with a uniform mesh of 1600 nodes are depicted over Fig. 4 where the computed and measured free surface elevations at four points are presented. Notice that for $\varepsilon = 10^{-7}$ m^{-2}.s^2 that is the most physical choice, the simulations of the complete and relaxed model cannot be distinguished.

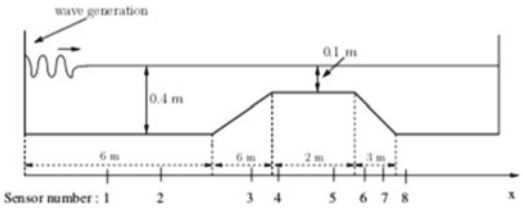


Fig. 3 Channel profile for the experiments and location of the sensors

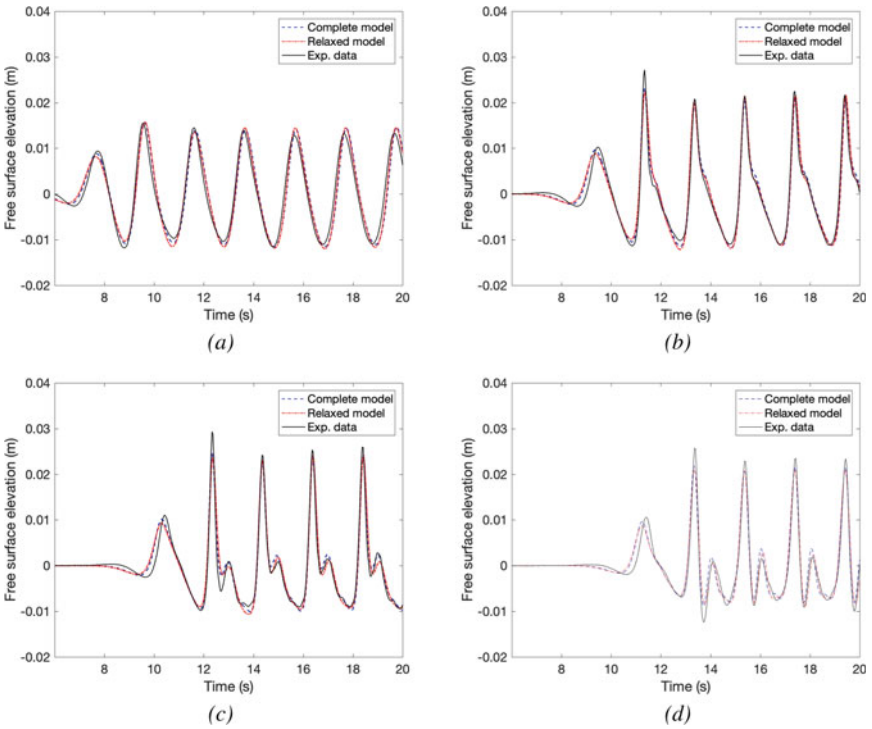


Fig. 4 Comparisons between the experimental data (solid line), the simulations of the dispersive model with the model presented in [2] (blue dashed line) and the relaxed model presented in this paper (red dashed-dotted line) with $\varepsilon = 10^{-4}$ m^{-2}.s^2. Figures *(a)*, *(b)*, *(c)* and *(d)* respectively correspond to the results for the sensors 3, 4, 5 and 6

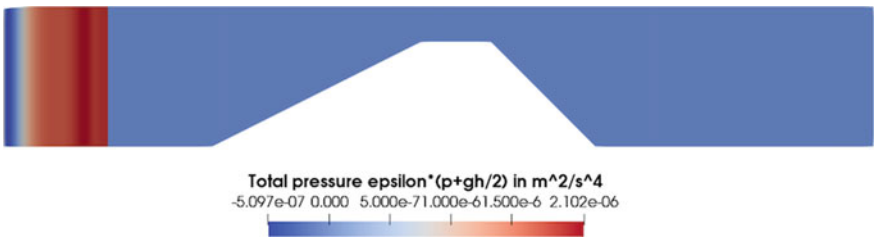


Fig. 5 Variations of the quantity $x \mapsto \varepsilon(p + gh/2)$ with $\varepsilon = 10^{-7}$ m^{-2}.s^2 at time $t = 0.01$ s

For the test case depicted in Sect. 4.5.1, the basin is at rest at the initial instant and we give at time $t = 0.01$ s, the value of the quantity $\varepsilon(p + gh/2)$ representing the pseudo-compressible effects. It appears over Fig. 5 that at time $t = 0.1$ s, whereas the free surface has just begun to deform at the boundary where the wave is generated, the acoustic-type waves have already propagated in the basin.

4.5.2 Comparison of the Computational Costs

For the simulation results given in Sect. 4.5.1, we compare the computational costs of the numerical schemes with and without pseudo-compressibility effects. More precisely, we compare the CPU time necessary to simulate the test case presented in Sect. 4.5.1 with the method proposed in [2]—corresponding to an incompressible model and requiring to solve the elliptic Eq. (127)—and the proposed explicit in time scheme (122)–(123) with the pseudo-compressible effects.

The advantages of the model and numerical strategy presented in this paper are significant for 2d problems with a large number of nodes but can hardly be highlighted in the 1d case where the elliptic operator to inverse is a symmetric tridagonal matrix. Hence, in order to illustrate the interest of the proposed scheme, we have used a conjugate gradient technique mimicking what would be done to solve (127) in 2d for an unstructured mesh.

Figure 6 presents the CPU time required to perform the simulations of the Dinguemans experiment with several meshes namely with 2000, 4000, 8000, 16000 and 32000 nodes. It appears that when the number of nodes increases, the proposed explicit in time scheme is more efficient than the conjugate gradient algorithm (used here without preconditioning). Notice that the authors have not performed an exhaustive comparison between the costs of the conjugate gradient technique—for which several optimizations are possible—and the iterative and explicit time resolution scheme (122)–(123).

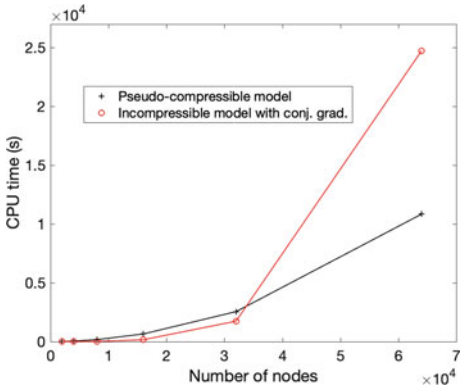


Fig. 6 Computational costs necessary to simulate the Dinguemans experiment with several meshes

Acknowledgements The authors thank François Bouchut for his helpful and constructive discussions that greatly contributed to improve the final version of the paper.

References

1. Airy, G.B.: Tides and waves. Encycl. Metropolitana **5**, 291–369 (1845)
2. Aïssiouene, N., Bristeau, M.-O., Godlewski, E., Mangeney, A., Parés, C., Sainte-Marie, J.: A two-dimensional method for a family of dispersive shallow water model. Working Paper or Preprint, May 2020. https://hal.archives-ouvertes.fr/hal-01632522
3. Audusse, E., Bouchut, F., Bristeau, M.-O., Klein, R., Perthame, B.: A fast and stable well-balanced scheme with hydrostatic reconstruction for Shallow Water flows. SIAM J. Sci. Comput. **25**(6), 2050–2065 (2004)
4. Audusse, E., Bouchut, F., Bristeau, M.-O., Sainte-Marie, J.: Kinetic entropy inequality and hydrostatic reconstruction scheme for the Saint-Venant system. Math. Comp. **85**(302), 2815–2837 (2016). MR 3522971
5. Audusse, E., Bristeau, M.-O.: Transport of pollutant in shallow water flows: a two time steps kinetic method. ESAIM: M2AN **37**(2), 389–416 (2003)
6. Audusse, E., Bristeau, M.-O.: A well-balanced positivity preserving second-order scheme for Shallow Water flows on unstructured meshes. J. Comput. Phys. **206**(1), 311–333 (2005)
7. Barré de Saint-Venant, A.-J.-C.: Théorie du mouvement non permanent des eaux avec applications aux crues des rivières et à l'introduction des marées dans leur lit. C. R. Acad. Sci. Paris **73**, 147–154 (1871)
8. Bona, J.-L., Benjamin, T.-B., Mahony, J.-J.: Model equations for long waves in nonlinear dispersive systems. Philos. Trans. Roy. Soc. London Ser. A **272**, 47–78 (1972)
9. Bona, J.L., Chen, M., Saut, J.-C.: Boussinesq equations and other systems for small-amplitude long waves in nonlinear dispersive media: part I. Derivation and linear theory. J. Nonlinear Sci. **12**, 283–318 (2002)
10. Bouchut, F.: Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources. Birkhäuser (2004)
11. Boyer, F.: Trace theorems and spatial continuity properties for the solutions of the transport equation. Differ. Integr. Eqn. **18**(8), 891–934 (2005)
12. Bristeau, M.-O., Mangeney, A., Sainte-Marie, J., Seguin, N.: An energy-consistent depth-averaged Euler system: derivation and properties. Discrete Continuous Dyn. Syst. - Ser. B **20**(4), 961–988 (2015)
13. Camassa, R., Holm, D.D., Levermore, C.D.: Long-time effects of bottom topography in shallow water. Phys. D **98**(2–4), 258–286 (1996). Nonlinear phenomena in ocean dynamics (Los Alamos, NM, 1995). MR 1422281 (98a:76005)
14. Chorin, A.J.: Numerical solution of the Navier-Stokes equations. Math. Comp. **22**, 745–762 (1968). MR 0242392 (39 #3723)
15. Dedner, A., Kemm, F., Kroner, D., Munz, C.-D., Schnitzer, T., Wesenberg, M.: Hyperbolic divergence cleaning for the MHD equations. J. Comput. Phys. **172**(2), 645–673 (2002)
16. Desjardins, B.: A few remarks on ordinary differential equations. Commun. Part. Differ. Eqn. **21**(11–12), 1667–1703 (1996)
17. Di Perna, R.J., Lions, P.L.: Ordinary differential equations, transport theory and Sobolev spaces. Invent. Math. **98**, 511–547 (1989)
18. Dingemans, M.-W.: Wave propagation over uneven bottoms. Advanced Series on Ocean Engineering. World Scientific (1997)
19. Duchêne, V.: Rigorous justification of the Favrie-Gavrilyuk approximation to the Serre-Green-Naghdi model. Nonlinearity **32**(10), 3772–3797 (2019)
20. Escalante, C., Dumbser, M., Castro, M.J.: An efficient hyperbolic relaxation system for dispersive non-hydrostatic water waves and its solution with high order discontinuous Galerkin schemes. J. Comput. Phys. **394**, 385–416 (2019)

21. Gerbeau, J.-F., Perthame, B.: Derivation of viscous Saint-Venant system for Laminar Shallow Water; Numerical validation. Discrete Contin. Dyn. Syst. Ser. B **1**(1), 89–102 (2001)
22. Godlewski, E., Raviart, P.-A.: Numerical Approximation of Hyperbolic Systems of Conservation Laws. Applied Mathematical Sciences, vol. 118. Springer, New York (1996)
23. Green, A.E., Naghdi, P.M.: A derivation of equations for wave propagation in water of variable depth. J. Fluid Mech. **78**, 237–246 (1976)
24. Guermond, J.-L., Popov, B., Tovar, E., Kees, C.: Robust explicit relaxation technique for solving the Green-Naghdi equations. J. Comp. Phys. **399**, 108917 (2019)
25. LeVeque, R.-J.: Finite Volume Methods for Hyperbolic Problems. Cambridge University Press, Cambridge (2002)
26. Lions, P.-L.: Sur les équations différentielles ordinaires et les équations de transport. Comptes Rendus l'Acad. Sci. - Ser. I - Math. **326**(7), 833–838 (1998)
27. Nwogu, O.: Alternative form of Boussinesq equations for nearshore wave propagation. J. Waterway Port Coast. Ocean Eng. ASCE **119**(6), 618–638 (1993)
28. Peregrine, D.H.: Long waves on a beach. J. Fluid Mech. **27**, 815–827 (1967)
29. Unesco: Thermodynamic equation of seawater: calculation and use of thermodynamic properties (2010). http://unesdoc.unesco.org/images/0018/001881/188170e.pdf

A Generalised Serre-Green-Naghdi Equations for Variable Rectangular Open Channel Hydraulics and Its Finite Volume Approximation

Mohamed Ali Debyaoui and Mehmet Ersoy

Abstract We present a non-linear dispersive shallow water model which enters in the framework of section-averaged models. These new equations are derived up to the second order of the shallow water approximation starting from the three-dimensional incompressible and irrotational Euler system. The derivation is carried out in the case of non-uniform rectangular section and it generalises the well-known one-dimensional Serre-Green-Naghdi (SGN) equations on uneven bottom. The section-averaged model is asymptotically consistent with the Euler system in terms of mass, momentum, and energy equation which provides the richness of content for this model. We propose a well-balanced finite volume approximation and we present some numerical results to show the influence of the section variation.

Keywords Open channel flow · Euler equations · Asymptotic approximation · Serre-Green-Naghdi equations · Free surface shallow water equations · Non-hydrostatic pressure · Dispersive model · Finite volume

1 Introduction

In environmental modeling of free surface flows, whenever the aspect-ratio of the domain is small enough, the shallow water approximation is introduced to obtain reduced model for which the computational cost is lower than the one implied by the numerical solution of the full three-dimensional free surface equations. One of the most widely used models to describe the channel and river motion of watercourses is the *section-averaged free surface model* [2, 7, 8] which is a generalisation of the well-known *Saint-Venant system* (introduced by Adhémar Jean Claude Barré de Saint-Venant in the 19th Century [18]):

M. A. Debyaoui (✉) · M. Ersoy
Université de Toulon, IMATH EA 2134, 83957 La Garde, France
e-mail: Mohamed-Ali-Debyaoui@etud.univ-tln.fr

M. Ersoy
e-mail: Mehmet.Ersoy@univ-tln.fr

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*,
SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2_11

$$\begin{cases} \partial_t A + \partial_x Q & = 0, \\ \partial_t Q + \partial_x \left(\dfrac{Q^2}{A} + I_1(x, A) \right) & = I_2(x, A). \end{cases} \tag{1}$$

In these equations, $A = \sigma h$ is the wet area of fluid cross-section, Q is the water discharge, $I_1(x, A) = \frac{A^2}{2F_r^2 \sigma}$ is the *hydrostatic pressure* where F_r is the Froude's number and $I_2(x, A) = \frac{\sigma'(x)}{\sigma(x)} \dfrac{A^2}{2{F_r}^2 \sigma(x)} - \dfrac{A}{{F_r}^2} d'(x)$ is the hydrostatic pressure source term which takes into account the variation of the channel width σ and the bottom d. The model (1) reduces to the well-known one-dimensional Saint-Venant equations for uniform rectangular section, i.e. if σ is constant. The free surface model is the first order shallow water approximation of the section-averaged Navier-Stokes or Euler equations under suitable assumptions on the horizontal and the vertical scales (see, e.g., [2, 7, 8, 10, 11] and the reference therein).

As it is well-known, the solutions of these equations are usually suitable to approximate breaking waves with turbulent rollers for large transitions of the Froude's number. However, for small or moderate transitions, the solutions of these equations are not able to catch undular bores induced by a non-hydrostatic pressure distribution [17]. Up to our knowledge, the first section-averaged dispersive shallow water equations for quite general assumptions on the geometry of the channel was proposed in [6], thus allowing for the application of the resulting equations to natural rivers with arbitrarily shaped cross-sections. This model reads

$$\begin{cases} \partial_t A + \partial_x Q = 0 \\ \partial_t Q + \partial_x \left(\dfrac{Q^2}{A} + I_1(x, A) + \mu_2 DI_1(x, A, Q) \right) = I_2(x, A) + \mu_2 DI_2(x, A, Q) + O(\mu_2^2) \end{cases}$$

where DI_1 and DI_2 are the non-hydrostatic counterparts of the hydrostatic pressure and the hydrostatic pressure source term. The case of non-uniform rectangular section can be regarded as the natural extension of the usual one-dimensional Serre-Green-Naghdi (SGN) equations over uneven bottom [5, 12, 19].

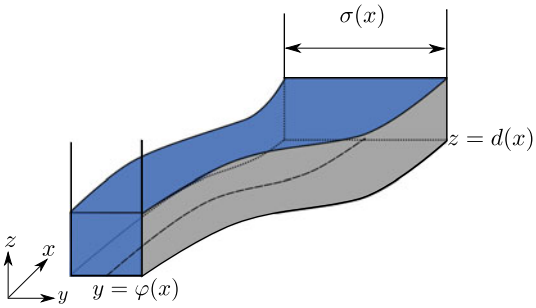
In this work, we focus only on the case of a rectangular variable section. We first present the geometrical set-up in Sect. 2. Then we give the outline of the asymptotic derivation, and in particular, we show that the section-averaged model is fully consistent with the Euler system in Sect. 3. Finally, in Sect. 4, we construct a first order well-balanced finite volume approximation and we present some numerical test cases.

2 The Three-Dimensional Incompressible Euler Equations

2.1 Settings

We consider the motion of an incompressible and irrotational fluid with constant density $\rho_0 > 0$ in a three-dimensional domain (see Fig. 1)

Fig. 1 Geometric set-up



$$\Omega(t) = \left\{(x, y, z) \in \mathbb{R}^3;\ x \in [0, L_c],\ \alpha(x) \leq y - \varphi(x) \leq \beta(x),\ d(x) \leq z \leq \eta(t, x, y)\right\}$$

where φ describes the transversal variation of the channel with respect to the main channel direction, α and β are the transversal limits of the channel, L_c its length, d is the bottom, η is the free surface and $h = \eta - d$ is the water height. The boundary of the domain $\Omega(t)$ is defined by $\partial\Omega(t)$ and is decomposed into four parts: the free surface $\Gamma_{\mathrm{fs}}(t)$, the wet boundary $\Gamma_{\mathrm{wb}}(t)$, the inflow boundary $\Gamma_{\mathrm{i}}(t)$ and the outflow boundary $\Gamma_{\mathrm{o}}(t)$. The wet boundary can be decomposed itself in three parts: the bottom $\Gamma_{\mathrm{b}}(t)$, the left lateral boundary $\Gamma_{\mathrm{lb}}(t)$, and the right one $\Gamma_{\mathrm{rb}}(t)$.

The governing equations for the motion of the fluid are the incompressible and irrotational Euler equations in $\Omega(t)$, for all $t \in (0, T]$, which can be written as follows:

$$\begin{aligned} \operatorname{div}[\boldsymbol{u}] &= 0\,, \\ \frac{\partial}{\partial t}(\boldsymbol{u}) + \operatorname{div}[\boldsymbol{u} \otimes \boldsymbol{u}] + \nabla\frac{p}{\rho_0} - \boldsymbol{F} &= 0 \end{aligned} \tag{2}$$

where $\boldsymbol{u} = (u, v, w)^T$ is the velocity field, $\boldsymbol{F} = (0, 0, -g)^T$ is the gravity acceleration and p is the pressure. These equations are completed by the irrotational equation:

$$\operatorname{curl}[\boldsymbol{u}] = 0. \tag{3}$$

The system is closed by suitable boundary conditions. We denote by $\boldsymbol{n}_{\mathrm{fs}}$ the outward normal to the free surface which depends on time:

$$\boldsymbol{n}_{\mathrm{fs}} = \frac{1}{\sqrt{1 + (\partial_x\eta)^2 + \left(\partial_y\eta\right)^2}}\ \left(-\partial_x\eta,\ -\partial_y\eta,\ 1\right)^T,$$

and by $\boldsymbol{n}_{\mathrm{wb}}$ the outward normal to the wet boundary:

$$\boldsymbol{n}_{\mathrm{wb}} = \begin{cases} \dfrac{1}{\sqrt{1+(\partial_x d)^2}}\ (\partial_x d,\ 0,\ -1)^T & \text{if } \boldsymbol{n}_{\mathrm{wb}} = \boldsymbol{n}_{\mathrm{b}} \\ \dfrac{1}{\sqrt{1+(\partial_x \alpha)^2}}\ (\partial_x \alpha,\ -1,\ 0)^T & \text{if } \boldsymbol{n}_{\mathrm{wb}} = \boldsymbol{n}_{\mathrm{lb}} \\ \dfrac{1}{\sqrt{1+(\partial_x \beta)^2}}\ (\partial_x \beta,\ 1,\ 0)^T & \text{if } \boldsymbol{n}_{\mathrm{wb}} = \boldsymbol{n}_{\mathrm{rb}} \end{cases}$$

At the free surface, we prescribe a kinematic boundary condition

$$\partial_t \eta + u\partial_x \eta + v\partial_y \eta = w \text{ on } \Gamma_{\mathrm{fs}}(t) \tag{4}$$

completed with the dynamical condition which takes into account the equilibrium with atmospheric stress

$$p = p_a \text{ on } \Gamma_{\mathrm{fs}}(t). \tag{5}$$

In the sequel, without loss of generality, we set $p_a = 0$.

At the wet boundary, we prescribe a no-penetration condition:

$$\begin{array}{l} u\partial_x d - w = 0 \ \text{ on } \ \Gamma_{\mathrm{b}}(t), \\ u\partial_x \alpha - v \ = 0 \ \text{ on } \ \Gamma_{\mathrm{lb}}(t), \\ u\partial_x \beta + v \ = 0 \ \text{ on } \ \Gamma_{\mathrm{rb}}(t). \end{array} \tag{6}$$

2.2 Dimensionless Euler Equations

Let us consider the following scales involved in the wave motion: L a wave-length in the longitudinal direction, H_2 a characteristic water depth, H_1 a characteristic scale of the channel width and h_1 a wave-length in the transversal direction. We then define the classical dispersive parameter μ_2 (see e.g. [13])

$$\mu_2 = \frac{H_2^2}{L^2}$$

and $\mu_1 = \frac{h_1^2}{L^2}$ where μ_1 is also a dispersive parameter but in the transversal direction.

In the following, we consider the asymptotic regime:

$$h_1 < H_1 = H_2 \ll L$$

such that the following inequality holds

$$\mu_1 < \mu_2^2.$$

Under these assumptions, we get the following ordering:

$$\mu_1^2 < \frac{\mu_1^2}{\mu_2} < \min\left(\frac{\mu_1^2}{\mu_2^2}, \mu_1\mu_2\right) < \max\left(\frac{\mu_1^2}{\mu_2^2}, \mu_1\mu_2\right) < \mu_1 < \min\left(\frac{\mu_1}{\mu_2}, \mu_2^2\right) < \max\left(\frac{\mu_1}{\mu_2}, \mu_2^2\right) < \mu_2 \ll 1.$$

We also introduce $(U, V = \sqrt{\mu_1}U, W = \sqrt{\mu_2}U)^T$ the scale of fluid velocity. The time scale is $T = \dfrac{L}{U}$. Let us define $P = \dfrac{p}{\rho_0}$ and choose the pressure scale to be $\mathcal{P} = U^2$.

This allows us to introduce the dimensionless quantities of time $\widetilde{t}$, space $(\widetilde{x}, \widetilde{y}, \widetilde{z})$, pressure $\widetilde{P}$, depth $\widetilde{d}$, water elevation $\widetilde{\eta}$ and velocity field $(\widetilde{u}, \widetilde{v}, \widetilde{w})$, via the following scaling relation

$$\widetilde{x} = \frac{x}{L}, \widetilde{y} = \frac{y}{h_1}, \widetilde{z} = \frac{z}{H_2}, \widetilde{t} = \frac{t}{T}, \widetilde{P} = \frac{P}{\mathcal{P}}, \widetilde{\varphi} = \frac{\varphi}{h_1}, \widetilde{u} = \frac{u}{U}, \widetilde{d} = \frac{d}{H_2}, \widetilde{v} = \frac{v}{V}, \widetilde{\eta} = \frac{\eta}{H_2}, \widetilde{w} = \frac{w}{W}. \tag{7}$$

Finally, we define the non-dimensional Froude's number by $F_r = \dfrac{U}{\sqrt{gH_2}}$.

For the sake of clarity and simplicity dropping $\widetilde{\cdot}$, using the dimensionless variables (7), and reordering the terms with respect to the powers of μ_1 and μ_2, the dimensionless incompressible Euler system (2) reads as follows:

$$\partial_x u + \partial_y v + \partial_z w = 0, \tag{8}$$

$$\partial_t u + u\partial_x u + v\partial_y u + w\partial_z u + \partial_x P = 0, \tag{9}$$

$$\mu_1\left(\partial_t v + u\partial_x v + v\partial_y v + w\partial_z v\right) + \partial_y P = 0, \tag{10}$$

$$\mu_2\left(\partial_t w + u\partial_x w + v\partial_y w + w\partial_z w\right) + \partial_z P = -\frac{1}{{F_r}^2}. \tag{11}$$

Under this scaling, the boundary conditions (4)–(5) and (6) remain unchanged and the dimensionless irrotational Eq. (3) becomes

$$\partial_y u = \mu_1\partial_x v, \ \mu_1\partial_z v = \mu_2\partial_y w, \ \partial_z u = \mu_2\partial_x w. \tag{12}$$

Thanks to the ordering $\mu_1^2 < \mu_2$ and the structure of Eqs. (12), it is natural to compute the asymptotic expansion of u in two steps first with respect to y, then with respect to z. It can be achieved by first width-averaging the Euler system (8)–(11), then by depth-averaging the resulting equations. For the sake of completeness, skipping the technical details, we present the outline of the derivation. Interested readers can found the details in [6].

3 Derivation of the Section-Averaged Model

3.1 Width-Averaged Equations

By integrating for $s \in [\alpha(x), y]$, the first two equations of the irrotational equations (12) and the divergence equation (8), keeping in mind the boundary conditions (4)–(5) and (6), we get the following asymptotic expansions:

$$u(t,x,y,z) = u_\alpha(t,x,z) - \frac{\mu_1}{2}\partial_x \mathrm{div}_{x,z}\left[\boldsymbol{w}_\alpha(t,x,z)(y-\alpha(x))^2\right] + O\left(\frac{\mu_1^2}{\mu_2}\right), \tag{13}$$

$$v(t,x,y,z) = -\mathrm{div}_{x,z}\left[\boldsymbol{w}_\alpha(t,x,z)(y-\alpha(x))\right] + O\left(\frac{\mu_1}{\mu_2}\right) \tag{14}$$

and

$$w(t,x,y,z) = w_\alpha(t,x,z) - \frac{\mu_1}{2\mu_2}\partial_z \mathrm{div}_{x,z}\left[\boldsymbol{w}_\alpha(t,x,z)(y-\alpha(x))^2\right] + O\left(\frac{\mu_1^2}{\mu_2^2}\right) \tag{15}$$

where $X_\alpha(t,x,z) := X(t,x,\alpha(x),z)$.

For a given function $(t,x,y,z) \mapsto X(t,x,y,z)$, we define its width-average by

$$\langle X\rangle(t,x,z) := \frac{1}{\sigma(x)}\int_{\alpha(x)}^{\beta(x)} X(t,x,y,z)\,dy$$

where $\sigma(x) = \beta(x) - \alpha(x)$ is the width of the channel.

Integrating Eqs. (8)–(11) for $y \in [\alpha(x), \beta(x)]$, using Leibniz integral rule, keeping in mind the boundary conditions (4)–(5) and (6), using the asymptotic expansions (13)–(15), we obtain the width-averaged Euler system:

$$\begin{aligned}
&\mathrm{div}_{x,z}\left[\sigma \boldsymbol{w}_\alpha\right] && = O\left(\frac{\mu_1}{\mu_2}\right),\\
&\frac{\partial}{\partial t}(\sigma u_\alpha) + \mathrm{div}_{x,z}\left[\sigma u_\alpha \boldsymbol{w}_\alpha\right] + \frac{\partial}{\partial x}(\sigma P_\alpha) + && = P_\alpha\frac{\partial \sigma}{\partial x} + O\left(\frac{\mu_1}{\mu_2}\right),\\
&\mu_2\left(\frac{\partial}{\partial t}(\sigma w_\alpha) + \mathrm{div}_{x,z}\left[\sigma w_\alpha \boldsymbol{w}_\alpha\right]\right) + \frac{\partial}{\partial z}(\sigma P_\alpha) && = -\frac{\sigma}{{F_r}^2} + P_\alpha\frac{\partial \sigma}{\partial z} + O(\mu_1)
\end{aligned} \tag{16}$$

where $P_\alpha(t,x,z) + O(\mu_1) = P(t,x,y,z)$ thanks to Eq. (10). The motion of the fluid is now in a two-dimensional domain:

$$\langle\Omega\rangle(t) = \left\{(x,z) \in \mathbb{R};\ d(x) \le z \le \eta^*(t,x)\right\}.$$

The irrotational condition (12) reduces to

$$\frac{\partial u_\alpha}{\partial z} = \mu_2 \frac{\partial w_\alpha}{\partial x} + O(\mu_1) \tag{17}$$

and the boundary conditions to

$$\frac{\partial \eta^*}{\partial t} + u_\alpha \frac{\partial \eta^*}{\partial x} = w_\alpha + O\left(\frac{\mu_1}{\mu_2}\right) \text{ and } P_\alpha = O(\mu_1) \text{ on } \langle \Gamma_{\text{fs}} \rangle(t) \,, \tag{18}$$

$$u_\alpha \partial_x d = w_\alpha + O\left(\frac{\mu_1}{\mu_2}\right) \text{ on } \langle \Gamma_{\text{b}} \rangle(t) \tag{19}$$

where $\langle \Gamma_{\text{fs}} \rangle(t)$ is the free surface boundary and $\langle \Gamma_{\text{b}} \rangle(t)$ the bottom boundary of the width-averaged fluid domain $\langle \Omega \rangle(t)$.

The function η^* in the above expression depends only on t and x. Indeed, integrating Eq. (11) for $s \in [z, \eta(t, x, y)]$, using the previous asymptotic expansions, and noting $\dfrac{D}{Dt} w = \partial_t w + u \partial_x w + v \partial_y w + w \partial_z w$, we can write

$$P_\alpha(t, x, z) = \frac{\eta(t, x, y) - z}{{F_r}^2} + \mu_2 \int_z^{\eta(t,x,y)} \frac{D}{Dt} w_\alpha(t, x, s)\, ds + O(\mu_1) \,.$$

Thus, taking the y-derivative of the above expression provides

$$0 = \partial_y \eta \left(\frac{1}{F_r^2} + \mu_2 \frac{D}{Dt} w_{\alpha|z=\eta} \right) + O(\mu_1) = -\partial_y \eta \, \partial_z P_{|z=\eta} + O(\mu_1)$$

Consequently, since $\partial_z P_{|z=\eta} \neq 0$, we get $\partial_y \eta = O(\mu_1)$. This is the so-called *flat free surface approximation*. Therefore, one can write

$$\eta(t, x, y) = \eta^*(t, x) + O(\mu_1) \tag{20}$$

where the * is dropped in the following.

3.2 Depth-Averaged Equations

Integrating Eq. (17) together with the first equation of System (16) for $s \in [d(x), z]$, keeping in mind Eqs. (18)–(19), we obtain

$$u_\alpha(t, x, z) = u_d(t, x) - \mu_2 \int_{d(x)}^{z} \partial_x \mathcal{S}(u_d, x, s)\, ds + O(\mu_2^2)$$

and

$$w_\alpha(t,x,z) = -\frac{1}{\sigma(x)}\frac{\partial}{\partial x}\left(u_d(t,x)S(x,z)\right) + O(\mu_2)$$

where $\mathcal{S}(u,x,z) = \frac{1}{\sigma(x)}\frac{\partial}{\partial x}\left(uS(x,z)\right)$, $S(x,z) = \sigma(x)(z-d(x))$ and $X_d(t,x) = X_\alpha(t,x,d(x))$.

Thanks to the flat free surface approximation (20), one can write the section-average of the velocity u as follows:

$$\overline{u} = \frac{1}{A}\int_{d(x)}^{\eta(t,x)}\int_{\alpha(x)}^{\beta(x)} u(t,x,y,z)\,dy\,dz$$

where $A = \int_{d(x)}^{\eta(t,x)}\int_{\alpha(x)}^{\beta(x)} dy\,dz = \sigma(x)h(t,x)$ is the wet area, $\sigma = \beta - \alpha$ is the width of the channel and $h = \eta - d$ is the water height.

Thus, since $u(t,x,y,z) = u_\alpha(t,x,z) + O(\mu_1) = u_d(t,x) - \mu_2\int_{d(x)}^{z}\partial_x\mathcal{S}(u_d,x,s)ds + O(\mu_2^2)$, we deduce the following asymptotic expansion of u:

$$u = \overline{u}(t,x) + \mu_2 B_0(\overline{u},x,z) + O(\mu_2^2) \tag{21}$$

where

$$B_0(\overline{u},x,z) = \frac{1}{A(t,x)}\int_{d(x)}^{\eta(t,x)}\left(\sigma(x)\int_{d(x)}^{z}\partial_x\mathcal{S}(\overline{u},x,s)\,ds\right)dz - \int_{d(x)}^{z}\partial_x\mathcal{S}(\overline{u},x,s)\,ds.$$

Similarly, we get for w:

$$w(t,x,y,z) = -\mathcal{S}(\overline{u},x,z) + O\left(\frac{\mu_1}{\mu_2}\right). \tag{22}$$

Using the asymptotic expansion of u (21) and w (22) , we obtain the asymptotic expansion of the pressure P at order $O(\mu_2^2)$

$$P(t,x,y,z) = P_\alpha(t,x,z) + O(\mu_1) = P_{\mathrm{h}}(t,x,z) + \mu_2 P_{\mathrm{nh}}(t,x,z) + O(\mu_2^2)$$

where

$$P_{\mathrm{h}}(t,x,z) = \frac{(\eta(t,x)-z)}{{F_r}^2}$$

is the usual hydrostatic pressure and

$$\begin{aligned} P_{\mathrm{nh}}(t,x,z) = {} & \int_z^{\eta(t,x)}\frac{1}{2\sigma(x)^2}\partial_z\left(\left(\sigma(x)\mathcal{S}(\overline{u},x,s)\right)^2\right)ds \\ & - \int_z^{\eta(t,x)}\partial_t\mathcal{S}(\overline{u},x,s) + \tfrac{\overline{u}(t,x)}{\sigma(x)}\partial_x\left(\sigma(x)\mathcal{S}(\overline{u},x,s)\right)ds \end{aligned}$$

is the non-hydrostatic part of the pressure.

3.3 Section-Averaged Model

To end the asymptotic derivation, we integrate vertically the set of equations (16) between d and η and drop all terms of order lower than μ_2. We get the generalised Serre-Green-Naghdi equations for non-uniform rectangular section:

$$\begin{cases} \partial_t A + \partial_x Q = 0 \\ \partial_t Q + \partial_x \left(\dfrac{Q^2}{A} + I_1(x, A) + \mu_2 DI_1(x, A, Q) \right) = I_2(x, A) + \mu_2 DI_2 + O(\mu_2^2) \end{cases} \tag{23}$$

where $A = \sigma h$ is the wet area, Q is the water discharge, $I_1(x, A) = \frac{A^2}{2F_r{}^2\sigma(x)}$ is the hydrostatic pressure, $I_2(x, A) = \frac{\sigma'(x)}{\sigma(x)} \dfrac{A^2}{2F_r{}^2\sigma(x)} - \dfrac{A}{F_r{}^2} d'(x)$ is the hydrostatic pressure source term, $DI_1 = \int_{d(x)}^{\eta(t,x)} P_{\text{nh}}(t, x, z)\sigma(x)\, dz$ is the non-hydrostatic pressure and $DI_2 = \int_{d(x)}^{\eta(t,x)} P_{\text{nh}}(t, x, z)\sigma'(x)\, dz - \sigma(x) P_{\text{nh}}(t, x, d(x))d'(x)$ is the non-hydrostatic pressure source term.

Moreover, Eqs. (23) are by construction asymptotically consistent with the Euler system (8)–(11). We have the following result:

Theorem 1 *System (23) admits a total energy*

$$E = A\frac{\overline{u}^2}{2} + A\frac{\eta}{F_r^2} - I_1 + \frac{\mu_2}{2} \int_\Omega \mathcal{S}^2(\overline{u}, x, z)\, dydz \tag{24}$$

which satisfies the following energy equation

$$\partial_t E + \partial_x((E + I_1 + \mu_2 DI_1)\overline{u}) = 0. \tag{25}$$

Moreover, the quantity E is consistent with the total energy $\mathcal{E} = \frac{u^2+\mu_1 v^2+\mu_2 w^2}{2} + \frac{z}{F_r^2}$ of the Euler equation (8)–(11), in the sense that

$$\partial_t \int_\Omega \mathcal{E}\, dydz + \partial_x \int_\Omega (\mathcal{E} + P)u\, dydz = \partial_t E + \partial_x((E + I_1 + \mu_2 DI_1)\overline{u}) + O(\mu_2^2).$$

Remark 1 This is a positive feature of the approximate model (23), which provides the richness of content for this model and can be used in the estimation of the accuracy of numerical algorithms. Moreover, it is well-known that the energy conservation law plays a fundamental role in the justification of the theory of shallow water equations.

Remark 2 As a direct consequence of (24) and (25), we are able to recover the energy conservation law of the usual models in the case of $\sigma \equiv 1$, i.e. $A = h$:

- if $\mu_2 = 0$, we recover the classical total energy of the Saint-Venant system, namely

$$E = \frac{h\overline{u}^2}{2} + \frac{h(h+2d)}{2F_r^2}.$$

- if $\mu_2 \neq 0$, we recover the classical total energy of the Serre-Green-Naghdi system (see for instance [9]), namely

$$E = \frac{h\overline{u}^2}{2} + \frac{h(h+2d)}{2F_r^2} + \mu_2 \left(\frac{h^3}{6} (\partial_x \overline{u})^2 - d' \frac{h^2}{2} \partial_x \overline{u} + \frac{(d')^2}{2} \right).$$

4 A Well-Balanced Finite Volume Approximation

The main drawback of Eqs. (23) is that it has third order terms in space which may lead to instabilities at the numerical level. Therefore, we first propose a more stable formulation of Eqs. (23) before presenting its numerical approximation.

Skipping the technical details, defining a linear operator $\mathbb{L}$ (where $\mathcal{L}$ is defined below)

$$\mathbb{L}[A,d,\sigma](u) = A\mathcal{L}[A,d,\sigma]\left(\frac{u}{A}\right),$$

one can show that System (23) can be written:

$$\begin{cases} \partial_t A + \partial_x Q = 0 \\ (I_d - \mu_2 \mathbb{L}[A,d,\sigma]) \left(\partial_t (A\overline{u}) + \partial_x \left(\frac{Q^2}{A} \right) \right) + \partial_x I_1(x,A) + \mu_2 A Q[A,d,\sigma]\left(\frac{Q}{A}\right) = I_2(x,A) + O(\mu_2^2) \end{cases} \tag{26}$$

where $Q = A\overline{u}$ is the discharge, I_d is the identity operator, $\mathcal{L}$ is a linear operator

$$\begin{aligned} \mathcal{L}[A,d,\sigma](u) = & \frac{1}{A}\left[\partial_x\left(\overline{\mathcal{T}}[A,d,\sigma](u,\sigma)\right) - \overline{\mathcal{T}}[A,d,\sigma](u,\partial_x\sigma)\right] \\ & + \frac{1}{A}\sigma(x)d'(x)\,\mathcal{T}[A,d,\sigma,z=d(x)](u) \end{aligned}$$

and Q is a quadratic operator

$$\begin{aligned} Q[A,d,\sigma](u) = & \frac{1}{A}\left[\partial_x\left(\overline{\mathcal{G}}[A,d,\sigma](u,\sigma)\right) - \overline{\mathcal{G}}[A,d,\sigma](u,\partial_x\sigma)\right] \\ & + \frac{1}{A}\sigma(x)d'(x)\,\mathcal{G}[A,d,\sigma,z=d(x)](u) \end{aligned}$$

with $\mathcal{T}$, $\mathcal{G}$ are given by

$$\mathcal{T}[A,d,\sigma,z](u) = \partial_x(u)\int_z^\eta \frac{S(x,s)}{\sigma(x)}\,ds + u\int_z^\eta \frac{1}{\sigma(x)}\partial_x S(x,s)\,ds,$$

and

$$\mathcal{G}[A,d,\sigma,z](u) = \int_z^\eta 2(\partial_x u)^2 \frac{S(x,s)}{\sigma(x)} + \frac{u^2}{\sigma(x)}\left(\frac{\partial_x S(x,s)\partial_x\sigma(x)}{\sigma(x)} - \partial_x\partial_x S(x,s)\right) + \partial_x\left(\frac{u^2}{2}\right)\frac{S(x,s)\partial_x\sigma(x)}{\sigma(x)^2}\,ds$$

with

$$\overline{\mathcal{X}}[A,d,\sigma](u,\psi)=\int_{d(x)}^{\eta}\psi\mathcal{X}[A,d,\sigma,z](u)\,dz.$$

In particular, one can explicitly compute those operators:

- if $\sigma\in\mathbb{R}_*^+$ and $d\in\mathbb{R}$ are constant then we recover the standard one-dimensional SGN equations (see for instance [14–16]) over flat bottom with

$$\mathcal{L}[A,d,\sigma](u)=\mathcal{L}_0[A,\sigma](u)=\frac{1}{\sigma h}\partial_x\left(\frac{\sigma h^3}{3}\partial_x u\right)$$

and

$$\mathcal{Q}[A,d,\sigma](u)=\mathcal{Q}_0[A,\sigma](u)=\frac{1}{\sigma h}\partial_x\left(\frac{2}{3}\sigma h^3\left(\partial_x u\right)^2\right).$$

- if $\sigma\in\mathbb{R}_*^+$ is constant and $d=d(x)$ then we recover the standard one-dimensional SGN equations (see for instance [14–16]) over uneven bottom with

$$\mathcal{L}[A,d,\sigma](u)=\mathcal{L}_1[A,d,\sigma](u)=\mathcal{L}_0[A,\sigma](u)-\frac{1}{\sigma h}\partial_x\left(\frac{\sigma h^2}{2}ud'(x)\right)+\frac{h}{2}\partial_x ud'(x)-u\left(d'(x)\right)^2$$

and

$$\mathcal{Q}[A,d,\sigma](u)=\mathcal{Q}_1[A,d](u)=\mathcal{Q}_0[A,\sigma](u)+\frac{1}{\sigma h}\partial_x\left(\sigma\frac{h^2}{2}u^2d''(x)\right)+h\left(\partial_x u\right)^2d'(x)+u^2d'(x)\,d''(x).$$

- if $\sigma=\sigma(x)$ and $d=d(x)$ then we get the generalised one-dimensional SGN equations for non-uniform rectangular channel over uneven bottom with

$$\mathcal{L}[A,d,\sigma](u)=\mathcal{L}_1[A,d,\sigma](u)+\frac{1}{\sigma h}\partial_x\left(\sigma'(x)\frac{h^3}{3}u\right)-\frac{\sigma'(x)}{\sigma}\left(\partial_x u\frac{h^2}{3}+u\frac{h^2}{3}\frac{\sigma'(x)}{\sigma}-u\frac{h}{2}d'(x)\right)$$

and

$$\begin{aligned}\mathcal{Q}[A,d,\sigma](u)\;=\;&\mathcal{Q}_1[A,d,\sigma](u)+\frac{1}{\sigma h}\partial_x\left(\left(\sigma'(x)\right)^2\frac{u^2}{\sigma}\frac{h^3}{3}\right)+\frac{1}{\sigma h}\partial_x\left(d'(x)\sigma'(x)u^2\frac{h^2}{2}\right)\\&-\frac{1}{\sigma h}\partial_x\left(\sigma'(x)u^2\frac{h^3}{3}\right)+\partial_x\left(\partial_x\left(\frac{u^2}{2}\right)\sigma'(x)\frac{h^3}{3}\right)-\frac{1}{\sigma h}\sigma'(x)\mathcal{R}[A,d,\sigma](u)\end{aligned}$$

with

$$\begin{aligned}\mathcal{R}[A,d,\sigma](u)\;=\;&\left(\partial_x u\right)^2\frac{h^3}{3}+u^2\left(\frac{\sigma'(x)}{\sigma}\right)^2\frac{h^3}{3}+u^2\left(\frac{\sigma'(x)}{\sigma}\right)d'(x)\frac{h^2}{2}-u^2\left(\frac{\sigma''(x)}{\sigma}\right)^2\frac{h^3}{3}+u^2d''(x)\frac{h^2}{2}\\&+\partial_x\left(\frac{u^2}{2}\right)\frac{\sigma'(x)}{\sigma}\frac{h^3}{3}-u^2d'(x)\frac{\sigma'(x)}{\sigma}\frac{h^2}{2}-u^2\sigma'(x)\left(d'(x)\right)^2h+u^2\sigma''(x)d'(x)\frac{h^2}{2}\\&-\partial_x\left(\frac{u^2}{2}\right)\sigma'(x)d'(x)\frac{h^2}{2}.\end{aligned}$$

It is known that third order derivatives involved in the initial model (23) may create high frequencies instabilities, but the presence of the $\left(I_d-\mu_2\mathbb{L}[A,d,\sigma]\right)^{-1}$ in the second equation of (26) stabilises the equations with respect to these perturbations. Therefore, in the following, we construct a numerical scheme for Eqs. (26) instead of Eqs. (23).

4.1 Numerical Method

This section is devoted to the numerical method to solve the reformulated dispersive model (26). It is rather natural to split the hyperbolic part to the dispersive one as done by several authors (see for instance [3–5]).

Let $N \in \mathbb{N}^*$. Let us consider the following uniform mesh on $[0, L_c]$. Cells are denoted for every $i \in [0, N+1]$, by $m_i = (x_{i-1/2}, x_{i+1/2})$ with $x_i = \frac{x_{i-1/2}+x_{i+1/2}}{2}$ the cell center and $\delta x = x_{i+1/2} - x_{i-1/2}$ the space mesh. The interfaces $x_{1/2} = 0$ and $x = x_{N+1/2}$ denote the upstream and the downstream ends. We also consider a time discretisation t_n defined by $t_{n+1} = t_n + \delta t_n$ where the time step δt_n is computed through a CFL condition related to the hyperbolic part.

Let us first highlight that the still water steady state for Eqs. (26) is independent of μ_2. Indeed, one has $\forall \mu_2 > 0$, the still water steady state equation reads

$$u = 0, \frac{A}{\sigma} + d = h_0$$

for some positive h_0. As a consequence, the construction of a well-balanced scheme can be easily achieved considering only the hyperbolic part of Eqs. (26), for instance, by the use of the hydrostatic reconstruction (see for instance [1]).

Let us define $d_{i+1/2} = \max(d_i, d_{i+1})$ where $d_i = \frac{1}{\delta x}\int_{m_i} d(x)dx$, $\sigma_{i+1/2} = \max(\sigma_i, \sigma_{i+1})$ where $\sigma_i = \frac{1}{\delta x}\int_{m_i} \sigma(x)dx$ and let us define the reconstructed states

$$A^-_{i+1/2} = \sigma_{i+1/2}\left(\frac{A_i}{\sigma_i} + d_i - d_{i+1/2}\right),\ A^+_{i+1/2} = \sigma_{i+1/2}\left(\frac{A_{i+1}}{\sigma_{i+1}} + d_{i+1} - d_{i+1/2}\right)$$

with

$$U^-_{i+1/2} = (A^-_{i+1/2}, A^-_{i+1/2}u_i),\ U^+_{i+1/2} = (A^+_{i+1/2}, A^+_{i+1/2}u_{i+1})$$

where $U_i = (A_i, A_iu_i)^T \approx \frac{1}{\delta x}\int_{m_i}(A, Au)^T dx$.

Let us introduce the flux

$$F_1(U) = Q,\ F_2(U) = Q^2/A \text{ and } F_3(x, U) = I_1(x, A) + \mu_2\overline{\mathcal{G}}[A, d, \sigma](u, \sigma)$$

and

$$\mathfrak{S}(x, U) = I_2 + \mu_2\overline{\mathcal{G}}[A, d, \sigma](u, \partial_x\sigma) - \mu_2\sigma(x)d'(x)\,\mathcal{G}[A, d, \sigma, z = d(x)](u).$$

Then, one can write System (26) as follows:

$$\begin{aligned} &\partial_t A + \partial_x F_1(U) &&= 0\\ &\left(I_d - \mu_2\mathbb{L}[A, d, \sigma]\right)(\partial_t Q + \partial_x F_2(U)) + \partial_x F_3(x, U) - \mathfrak{S}(x, U) &&= 0 \end{aligned}$$

With these settings, we define the following numerical scheme:

$$
\begin{aligned}
A_i^{n+1} &= A_i^n - \tfrac{\delta t_n}{\delta x}\left(\mathcal{F}_1\left(U_{i+1/2}^{-,n}, U_{i+1/2}^{+,n}\right) - \mathcal{F}_1\left(U_{i-1/2}^{-,n}, U_{i-1/2}^{+,n}\right)\right)\\
Q_i^{*} &= Q_i^n - \tfrac{\delta t_n}{\delta x}\left(\mathcal{F}_2\left(U_{i+1/2}^{-,n}, U_{i+1/2}^{+,n}\right) - \mathcal{F}_2\left(U_{i-1/2}^{-,n}, U_{i-1/2}^{+,n}\right)\right)\\
Q_i^{n+1} &= Q_i^{*} - \tfrac{\delta t_n}{\delta x}(Y^n)_i
\end{aligned}
$$

where

$$
\mathcal{A}^n Y^n = \left(\mathcal{F}_3\left(x_{i+1/2}, U_{i+1/2}^{-,n}, U_{i+1/2}^{+,n}\right) - \mathcal{F}_3\left(x_{i-1/2}, U_{i-1/2}^{-,n}, U_{i-1/2}^{+,n}\right) + \mu_2 \mathcal{N}_i^n\right)_{1\le i\le N}.
$$

The matrix $\mathcal{A}^n$ is the cell-centered approximation of the linear operator $\left(I_d - \mu_2 \mathbb{L}[A, d, \sigma]\right)$ and $\mathcal{N}_i^n$ is the cell-centered approximation of $-\overline{\mathcal{G}}[A, d, \sigma]\,(u, \partial_x \sigma) + \sigma(x) d'(x)\, \mathcal{G}[A, d, \sigma, z = d(x)]\,(u)$.

The numerical fluxes are defined by

$$
\begin{aligned}
\mathcal{F}_1\left(U_{i+1/2}^{-,n}, U_{i+1/2}^{+,n}\right) &= \tfrac{F_1\left(U_{i+1/2}^{-,n}\right)+F_1\left(U_{i+1/2}^{+,n}\right)}{2} - s_{i+1/2}^n (A_{i+1/2}^{+,n} - A_{i+1/2}^{-,n})\\
\mathcal{F}_2\left(U_{i+1/2}^{-,n}, U_{i+1/2}^{+,n}\right) &= \tfrac{F_2\left(U_{i+1/2}^{-,n}\right)+F_2\left(U_{i+1/2}^{+,n}\right)}{2} - s_{i+1/2}^n (Q_{i+1/2}^{+,n} - Q_{i+1/2}^{-,n})\\
\mathcal{F}_3\left(x_{i+1/2}, U_{i+1/2}^{-,n}, U_{i+1/2}^{+,n}\right) &= \tfrac{F_3\left(x_{i+1/2}, U_{i+1/2}^{-,n}\right)+F_3\left(x_{i+1/2}, U_{i+1/2}^{+,n}\right)}{2} + \left(\tfrac{{A_i^n}^2}{2\sigma_i F_r^2} - \tfrac{{A_{i+1/2}^{-,n}}^2}{2\sigma_{i+1/2} F_r^2}\right)\\
\mathcal{F}_3\left(x_{i+1/2}, U_{i+1/2}^{-,n}, U_{i+1/2}^{+,n}\right) &= \tfrac{F_3\left(x_{i+1/2}, U_{i+1/2}^{-,n}\right)+F_3\left(x_{i+1/2}, U_{i+1/2}^{+,n}\right)}{2} + \left(\tfrac{{A_{i+1}^n}^2}{2\sigma_{i+1} F_r^2} - \tfrac{{A_{i+1/2}^{+,n}}^2}{2\sigma_{i+1/2} F_r^2}\right)
\end{aligned}
$$

such that whenever $\mu_2 = 0$, we recover the classical numerical scheme[1] for the hyperbolic part

$$
U_i^{n+1} = U_i^n - \frac{\delta t_n}{\delta x}\left(\mathcal{F}(x_{i+1/2}, U_{i+1/2}^{-,n}, U_{i+1/2}^{+,n}) - \mathcal{F}(x_{i-1/2}, U_{i-1/2}^{-,n}, U_{i-1/2}^{+,n})\right)
$$

with $\mathcal{F}(x, U, V) = (\mathcal{F}_1(U, V), \mathcal{F}_2(U, V) + \mathcal{F}_3(x, U, V))$. In these expressions,

$$
s_{i+1/2} = \max_{j=1,2}\left|\lambda_j(x_{i+1/2}, U_{i+1/2}^{-,n})\right|, \left|\lambda_j(x_{i+1/2}, U_{i+1/2}^{+,n})\right|
$$

where $\lambda_j(x, U) = Q/A + (-1)^j \sqrt{\frac{A}{\sigma(x) F_r^2}}$, $j = 1, 2$ are the eigenvalues of the Jacobian matrix of $(F_1, F_2 + F_3)^T$.

The numerical scheme is consistent and stable under the CFL condition

$$
\max_{1\le i\le N}\left(\left|\lambda_1(x_i, U_i^n)\right|, \left|\lambda_2(x_i, U_i^n)\right|\right)\frac{\delta t_n}{\delta x} \le 1\,.
$$

[1]For the sake of simplicity and clarity, we have presented the finite volume method using the Rusanov solver but the method is not limited to this one.

4.2 Propagation of a Solitary Wave

In this section, we test the accuracy of the method and we show numerically the influence of the section variation in the case of the propagation of a solitary wave. For this purpose, we consider the exact solitary wave solutions of the Green-Naghdi equations in the one-dimensional setting over a flat bottom (see [15]), given in variables with dimensions, by

$$\eta(t,x) = a\,\mathrm{sech}^2(k(x-ct)), \quad u(t,x) = c\left(\frac{\eta(t,x)}{\eta(t,x)+z_0}\right) \text{ with } k = \frac{\sqrt{3a}}{2z_0\sqrt{z_0+a}} \text{ and } c = \sqrt{g(z_0+a)} \tag{27}$$

where z_0 is the depth of the fluid and a is the relative amplitude.

Accuracy

The propagation of the solitary wave (27) is initially centered at $x_0 = 10$ m with a relative amplitude $a = 0.2$ m over a constant water depth $z_0 = 2$ m. The computational domain is $L_c = 100$ m and it is discretized with N cells. The single solitary wave propagates from left to right. In this test, since the solitary wave is initially far from boundaries, the boundary conditions do not affect the computation, thus we choose to impose free boundary conditions at the downstream and upstream ends. The exact solution is computed in a channel of width $\sigma = 1$.

In what follows, we quantify the numerical accuracy of our numerical scheme by computing the numerical solution for this particular test case for an increasing number of cells N over a duration $T = 20$ s. Starting with $N = 100$ number of cells, we successively multiply the number of cells by two. For all n, we compare, in Fig. 2,

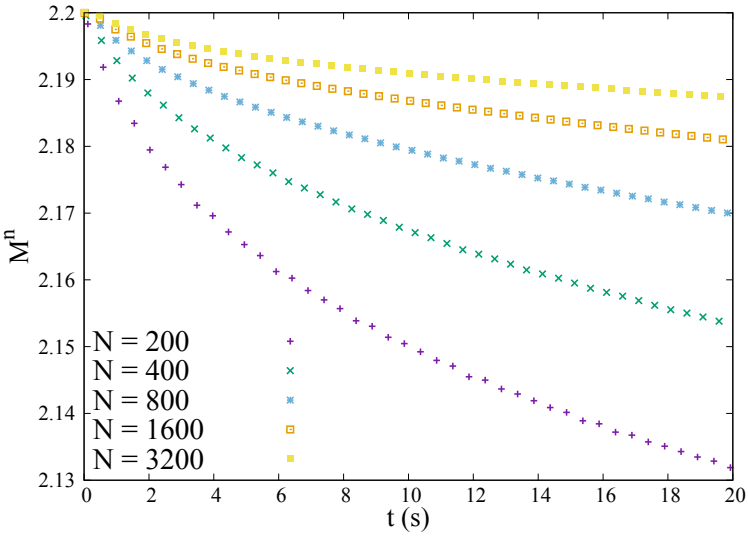


Fig. 2 $M^n := \max_{0\le i\le N+2}(h_i^n)$

$M^n := \max_{0\le i\le N+2}(h_i^n)$ of our numerical solution provided by Eqs. (26) with the exact one $M(t_n) := \max h(t_n, x)_{x\in[0,L_c]} = 2.2$ given by (27). One can easily remark that the first order discretisation is not accurate for long time simulation due to the numerical dissipation. However, to limit the numerical dissipation of the first order numerical scheme, one can either limit the simulation time or consider a very large number of cells. However, it is better to increase the order of the numerical scheme but this is left to future work. Therefore, in what follows, we consider a shorter simulation time and a large number of cells, just to illustrate the influence of the variation of the channel.

Influence of the Section Variation

We consider again the propagation of a solitary wave initially centered at $x_0 = 10$ m of relative amplitude $a = 0.2$ m, over a constant water depth $z_0 = 2$ m onto a computational domain of $L_c = 50$ m and discretized with $N = 5000$ cells. The final simulation time is $T = 8$ s. Initially starting with $(\eta(0, x), u(0, x))$ (see Eqs. (27)), we compute the numerical simulation for the channels defined by

$$\sigma(x;\varepsilon) = \beta(x;\varepsilon) - \alpha(x;\varepsilon) \text{ with } \beta = \frac{1}{2} - \frac{\varepsilon}{2}\exp\left(-\varepsilon^2\left(x - L/2\right)^2\right) \text{ and } \alpha = -\beta$$

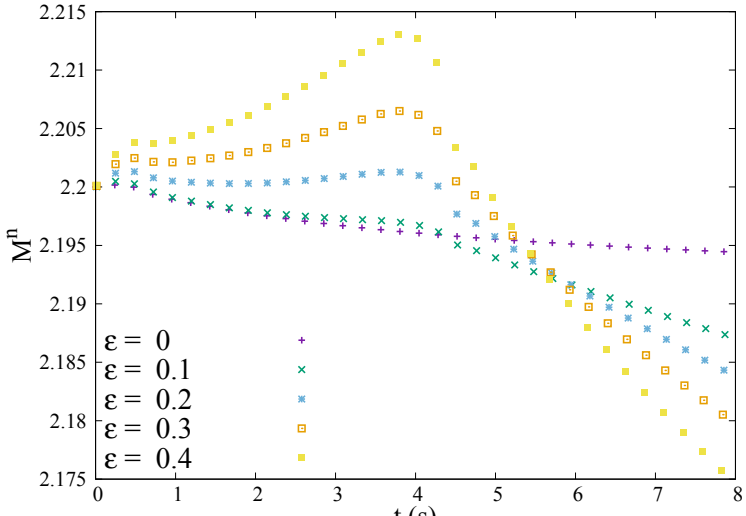
with $\epsilon = 0, \epsilon = 0.1, \epsilon = 0.2, \epsilon = 0.3$ and $\epsilon = 0.4$. The obtained results are presented in Fig. 3. In Fig. 3a, for each geometry, we show the evolution of the maximum of the water level $M^n := \max_{0\le i\le N+2}(h_i^n)$. As expected, since the first part for $x \le 25$ is linearly converging, the water level increases while for $x > 25$, the channel is linearly diverging and therefore, the amplitude of the water level decreases. Moreover, in all numerical simulations, the mass is conserved. Indeed, for each value of ε, we have displayed in Fig. 3b, the ratio of $\frac{m^n}{m^0}$ where $m^n := \frac{1}{N+2}\sum_{i=0}^{N+1} A_i^n$ is the mass of water at time t_n. The ratio $\frac{m^n}{m^0}$ is almost equal to 1, up to the order of accuracy of the numerical scheme.

In what follows, we quantify the numerical accuracy of our numerical scheme. Starting with $N = 100$ number of cells, we successively multiply the number of cells by two. The errors on the water surface deformation are presented in Table 1 and in Table 2. These errors are computed at $t = 8$ s using the L^2:

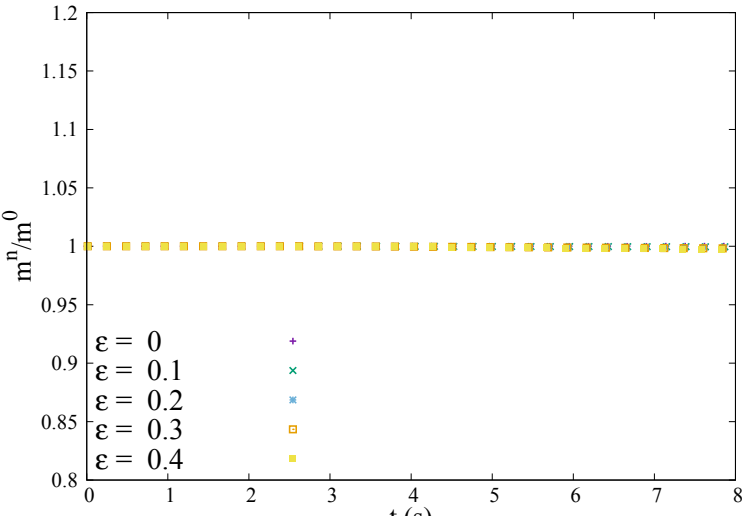
$$\|\eta_{\text{num}} - \eta_{\text{ref}}\|_2 = \sqrt{\delta x \sum_i \left|\eta_{\text{num}i}(t = 8) - \eta_{\text{ref}}(t = 8, x_i)\right|}$$

and the L^∞ norms where η_{ref} is the exact solution in the case $\varepsilon = 0$ and is a reference one computed with 10 000 cells for $\varepsilon = 0.4$. Since the results are almost the same whatever ε is, we have decided to present only the results for $\varepsilon = 0$ and $\varepsilon = 0.4$ in Table 1 and in Table 2.

As expected, the obtain numerical order is slow because of the numerical dissipation of the solitary wave (as already pointed out in several works, see for instance [3] for which we obtain almost the same order of convergence in the case of uniform section). Moreover, as expected, cross-sectional variations have no influence



(a) $M^n := \max_{x \in 0, L_c} h_i^n$



(b) $\frac{m^n}{m^0}$ with $m^n = \frac{1}{N+2} \sum_{i=0}^{N+1} A_i^n$

Fig. 3 Influence of σ

Table 1 Convergence rate of the L^2 error for $\varepsilon = 0$. The order is computed through first order interpolation polynomial

N	$\|\eta_{\text{num}} - \eta_{exact}\|_2$	$\|\eta_{\text{num}} - \eta_{exact}\|_\infty$
100	0.0789	0.0449
200	0.0497	0.0288
400	0.0304	0.0180
800	0.0198	0.0116
1600	0.0153	0.0081
3200	0.0138	0.0062
Order	0.53	0.58

Table 2 Convergence rate of the L^2 error for $\varepsilon = 0.4$. The reference solution is computed with 10 000 cells. The order is computed through first order interpolation polynomial

N	$\|\eta_{\text{num}} - \eta_{\text{ref}}\|_2$	$\|\eta_{\text{num}} - \eta_{\text{ref}}\|_\infty$
100	0.05212	0.02533
200	0.02096	0.01082
400	0.01079	0.00554
800	0.00748	0.00503
1600	0.00635	0.00412
3200	0.00505	0.00300
Order	0.64	0.56

on the convergence rate. Let us just emphasise that the convergence rates are slightly better in the case of non-uniform section because we are comparing our results to a reference solution and not to the exact one.

5 Conclusions and Perspectives

We have presented the derivation of a new dispersive model for open channel with non-uniform rectangular section. This model generalises the usual Serre-Green-Naghdi equation. We have presented its numerical finite volume approximation for which we have proposed two simple test cases. In a forthcoming paper, we will focus on the case of arbitrary channel section and we will propose a high order numerical scheme.

Acknowledgment The authors thank the referees for their valuable remarks which led to a substantial improvement of the first version of this paper. The second author wishes to grateful Dr Griggio for her help throughout this work and would like to say sincerely “Sağol”.

References

1. Audusse, E., Bouchut, F., Bristeau, M.O., Klein, R., Perthame, B.: A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. SIAM J. Sci. Comput. **25**, 2050–2065 (2004)
2. Bourdarias, C., Ersoy, M., Gerbi, S.: A mathematical model for unsteady mixed flows in closed water pipes. Sci. China Math. **55**, 221–244 (2012)
3. Bourdarias, C., Gerbi, S., Lteif, R.: A numerical scheme for an improved Green-Naghdi model in the Camassa-Holm regime for the propagation of internal waves. Comput. Fluids **156**, 283–304 (2017)
4. Chazel, F., Lannes, D., Marche, F.: Numerical simulation of strongly nonlinear and dispersive waves using a Green-Naghdi model. J. Sci. Comput. **48**, 105–116 (2011)
5. Cienfuegos, R., Barthélemy, E., Bonneton, P.: A fourth-order compact finite volume scheme for fully nonlinear and weakly dispersive boussinesq-type equations. Part II: boundary conditions and validation. Int. J. Numer. Methods. Fluids. **53**, 1423–1455 (2007)
6. Debyaoui, M.A., Ersoy, M.: Generalised Serre-Green-Naghdi equations for open channel and for natural river hydraulics (2020). https://hal.archives-ouvertes.fr/hal-02444355. Working paper or preprint
7. Decoene, A., Bonaventura, L., Miglio, E., Saleri, F.: Asymptotic derivation of the section-averaged shallow water equations for natural river hydraulics. Math. Models Methods Appl. Sci. **19**, 387–417 (2009)
8. Ersoy, M.: Dimension reduction for incompressible pipe and open channel flow including friction. In: Brandts, J., Korotov, S., Krizek, M., Segeth, K., Sistek, J., Vejchodsky, T. (eds.) Conference Applications of Mathematics 2015, in Honor of the 90th Birthday of Ivo Babuska and 85th Birthday of Milan Práger and Emil Vitásek , pp. 17–33, Prague, France (2015). https://hal.archives-ouvertes.fr/hal-00908961
9. Fedotova, Z.I., Khakimzyanov, G.S., Dutykh, D.: Energy equation for certain approximate models of long-wave hydrodynamics. Russ. J. Numer. Anal. Math. Model. **29**, 167–178 (2014)
10. Gerbeau, J.F., Perthame, B.: Derivation of viscous Saint-Venant system for laminar shallow water; numerical validation. Discrete Continuous Dyn. Syst. Ser. B **1**, 89–102 (2001)
11. Gouta, N., Maurel, F.: A finite volume solver for 1D shallow-water equations applied to an actual river. Int. J. Numer. Methods. Fluids. **38**, 1–19 (2002)
12. Green, A.E., Naghdi, P.M.: A derivation of equations for wave propagation in water of variable depth. J. Fluid Mech. **78**, 237–246 (1976)
13. Lannes, D.: The Water Waves Problem: Mathematical Analysis and Asymptotics, vol. 188. American Mathematical Society, Providence (2013)
14. Lannes, D., Bonneton, P.: Derivation of asymptotic two-dimensional time-dependent equations for surface water wave propagation. Phys. Fluids **21**, 016601 (2009)
15. Lannes, D., Marche, F.: A new class of fully nonlinear and weakly dispersive Green-Naghdi models for efficient 2D simulations. J. Comput. Phys. **282**, 238–268 (2015)
16. Lannes, D., Marche, F.: Nonlinear wave-current interactions in shallow water. Stud. Appl. Math. **136**, 382–423 (2016)
17. Peregrine, D.: Calculations of the development of an undular bore. J. Fluid Mech. **25**, 321–330 (1966)
18. de Saint-Venant, A.J.C.B.: Théorie du mouvement non-permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leur lit. C. R. Acad. Sci. **73**, 147–154 (1871)
19. Seabra-Santos, F.J., Renouard, D.P., Temperville, A.M.: Numerical and experimental study of the transformation of a solitary wave over a shelf or isolated obstacle. J. Fluid Mech. **176**, 117–134 (1987)

Author Index

M. L. Muñoz-Ruiz et al. (eds.), *Recent Advances in Numerical Methods for Hyperbolic PDE Systems*, SEMA SIMAI Springer Series 28, https://doi.org/10.1007/978-3-030-72850-2